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

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

continued

Vol. 19, No. 2—June 2025

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

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Continued from back cover

Neuroscience

Physical Science and Engineering

Social and Political Sciences

Sports

Vol. 19, No. 2—June 2025

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

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

continued

Medical and Health Sciences

Medical and Health Sciences: Deep Learning

Medical and Health Sciences: Electronic Health Records

Medical and Health Sciences: Epidemiology

Medical and Health Sciences: Mobile Health

Medical and Health Sciences: Reinforcement Learning

Continued on inner cover

Vol. 19, No. 2—June 2025

THE ANNALS OF APPLIED STATISTICS Vol. 19, No. 2, pp. 847–1799 June 2025

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

The Annals of Applied Statistics **[ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 19, Number 2, June 2025. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage pending at Cleveland, Ohio, and at additional mailing offices.

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

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 847–865
https://doi.org/10.1214/24-AOAS1920

SENSITIVITY ANALYSIS AND POWER IN THE PRESENCE OF MANY WEAK INSTRUMENTS: APPLICATION TO THE EFFECT OF INCARCERATION ON FUTURE EARNINGS

BY ASHKAN ERTEFAIE[1,a], JESSE Y. HSU[1,b], HARDING HARDING[2,c], JEFFREY MORENOFF[3,d] AND DYLAN S. SMALL[4,e]

[1]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania,* [a]*ertefaie@pennmedicine.upenn.edu,* [b]*Jesse.Hsu@pennmedicine.upenn.edu*

[2]*Department of Sociology, University of California at Berkeley,* [c]*dharding@berkeley.edu*

[3]*Department of Sociology, University of Michigan,* [d]*morenoff@umich.edu*

[4]*Department of Statistics, The Wharton School, University of Pennsylvania,* [e]*dsmall@wharton.upenn.edu*

This article discusses a sensitivity analysis for an instrumental variable (IV) estimate in the presence of many instruments that are weakly associated with the endogenous variable. We study the effect of imprisonment on earnings using data on individuals sentenced for felony in Michigan in the years 2003–2006. Motivated by the random assignment of judges to cases, we construct a vector of instruments based on judges' ID. Our data has two important features that cannot be handled using standard IV approaches. First, while some judges exhibit strong tendencies toward a prison or nonprison sentence, many judges do not have strong tendencies toward a particular sentence type. Second, our data includes only cases that result in sentencing, and thus the standard analyses are subject to selection bias. We develop a sensitivity analysis procedure that is robust to the presence of many weak instruments and quantifies the effect of the selection bias on the parameter of interest. A power formula for the sensitivity analysis is also provided. Analyses show that being sentenced to prison significantly reduces the offenders' earnings. Our simulation studies highlight the value of the proposed method in terms of statistical power and also confirm the validity of our power formula.

REFERENCES

ABADIE, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *J. Econometrics* **113** 231–263. MR1960380 https://doi.org/10.1016/S0304-4076(02)00201-4

ANDERSON, T. W., KUNITOMO, N. and MATSUSHITA, Y. (2010). On the asymptotic optimality of the LIML estimator with possibly many instruments. *J. Econometrics* **157** 191–204. MR2661594 https://doi.org/10.1016/j.jeconom.2009.12.001

ANDERSON, T. W. and RUBIN, H. (1949). Estimators of the parameters of a single equation in a complete set of stochastic equations. *Ann. Math. Stat.* **21**.

ANDREWS, I. and MIKUSHEVA, A. (2016). Conditional inference with a functional nuisance parameter. *Econometrica* **84** 1571–1612. MR3537164 https://doi.org/10.3982/ECTA12868

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

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

APEL, R. and SWEETEN, G. (2010). The impact of incarceration on employment during the transition to adulthood. *Soc. Probl.* **57** 448–479.

BAIOCCHI, M., CHENG, J. and SMALL, D. S. (2014). Instrumental variable methods for causal inference. *Stat. Med.* **33** 2297–2340. MR3257582 https://doi.org/10.1002/sim.6128

BEKKER, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* **62** 657–681. MR1281697 https://doi.org/10.2307/2951662

Key words and phrases. Instrumental variables, selection bias, test statistics, two-stage least squares, unmeasured confounders.

BLOOM, D., REDCROSS, C., ZWEIG, J. and AZURDIA, G. (2007). Transitional jobs for ex-prisoners: Early impacts from a random assignment evaluation of the center for employment opportunities (ceo) prisoner reentry program. Tech. Rep., Working Paper, New York, MDRC.

BOUND, J., JAEGER, D. A. and BAKER, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Amer. Statist. Assoc.* **90** 443–450.

BROOKHART, M. A., WANG, P., SOLOMON, D. H. and SCHNEEWEISS, S. (2006). Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* **17** 268.

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

CHAMBERLAIN, G. and IMBENS, G. (2004). Random effects estimators with many instrumental variables. *Econometrica* **72** 295–306. MR2031020 https://doi.org/10.1111/j.1468-0262.2004.00485.x

CHAO, J. C., HAUSMAN, J. A., NEWEY, W. K., SWANSON, N. R. and WOUTERSEN, T. (2014). Testing overidentifying restrictions with many instruments and heteroskedasticity. *J. Econometrics* **178** 15–21. MR3137889 https://doi.org/10.1016/j.jeconom.2013.08.003

CHAO, J. C. and SWANSON, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica* **73** 1673–1692. MR2156676 https://doi.org/10.1111/j.1468-0262.2005.00632.x

CHAO, J. C., SWANSON, N. R., HAUSMAN, J. A., NEWEY, W. K. and WOUTERSEN, T. (2012). Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments. *Econometric Theory* **28** 42–86. MR2899214 https://doi.org/10.1017/S0266466611000120

CHOI, I. and PHILLIPS, P. C. B. (1992). Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified structural equations. *J. Econometrics* **51** 113–150. MR1151956 https://doi.org/10.1016/0304-4076(92)90032-M

CONLEY, T. G., HANSEN, C. B. and ROSSI, P. E. (2012). Plausibly exogenous. *Rev. Econ. Stat.* **94** 260–272.

DAVIDSON, R. and MACKINNON, J. G. (1993). *Estimation and Inference in Econometrics*. The Clarendon Press, New York. MR1350531

DEMING, D. J. (2022). Four facts about human capital. *J. Econ. Perspect.* **36** 75–102.

DIPRETE, T. A. and GANGL, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociol. Method.* **34** 271–310.

ERTEFAIE, A., HSU, J. Y., HARDING, H., MORENOFF, J. and SMALL, D. S. (2025). Supplement to "Sensitivity analysis and power in the presence of many weak instruments: Application to the effect of incarceration on future earnings." https://doi.org/10.1214/24-AOAS1920SUPP

ERTEFAIE, A., NGUYEN, A., HARDING, D. J., MORENOFF, J. D. and YANG, W. (2018). Instrumental variable analysis with censored data in the presence of many weak instruments: Application to the effect of being sentenced to prison on time to employment. *Ann. Appl. Stat.* **12** 2647–2673. MR3875715 https://doi.org/10.1214/18-AOAS1174

ERTEFAIE, A., SMALL, D., FLORY, J. and HENNESSY, S. (2016a). Selection bias when using instrumental variable methods to compare two treatments but more than two treatments are available. *Int. J. Biostat.* **12** 219–232. MR3505695 https://doi.org/10.1515/ijb-2015-0006

ERTEFAIE, A., SMALL, D. S., FLORY, J. and HENNESSY, S. (2016b). A sensitivity analysis to assess bias due to selecting subjects based on treatment received. *Epidemiology* **27** e5–e7.

ERTEFAIE, A., SMALL, D. S., FLORY, J. H. and HENNESSY, S. (2017). A tutorial on the use of instrumental variables in pharmacoepidemiology. *Pharmacoepidemiol. Drug Saf.* **26** 357–367. https://doi.org/10.1002/pds.4158

ERTEFAIE, A., SMALL, D. S. and ROSENBAUM, P. R. (2018). Quantitative evaluation of the trade-off of strengthened instruments and sample size in observational studies. *J. Amer. Statist. Assoc.* **113** 1122–1134. MR3862344 https://doi.org/10.1080/01621459.2017.1305275

ESTELLE, S. M. and PHILLIPS, D. C. (2018). Smart sentencing guidelines: The effect of marginal policy changes on recidivism. *J. Public Econ.* **164** 270–293.

GOTTFREDSON, D. M. (1998). Choosing punishments: Crime control effects of sentences. Justice Policy Research Corporation.

HAAVELMO, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12. MR0007954 https://doi.org/10.2307/1905714

HAHN, J. and INOUE, A. (2002). A Monte Carlo comparison of various asymptotic approximations to the distribution of instrumental variables estimators. *Econometric Rev.* **21** 309–336. MR1944978 https://doi.org/10.1081/ETC-120015786

HANSEN, C., HAUSMAN, J. and NEWEY, W. (2008). Estimation with many instrumental variables. *J. Bus. Econom. Statist.* **26** 398–422. MR2459342 https://doi.org/10.1198/073500108000000024

HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123 https://doi.org/10.2307/1912775

HARDING, D. J., MORENOFF, J. D., NGUYEN, A. P. and BUSHWAY, S. D. (2017). Short-and long-term effects of imprisonment on future felony convictions and prison admissions. *Proc. Natl. Acad. Sci. USA* **114** 11103–11108.

HARDING, D. J., MORENOFF, J. D., NGUYEN, A. P. and BUSHWAY, S. D. (2018). Imprisonment and labor market outcomes: Evidence from a natural experiment. *Amer. J. Sociol.* **124** 49–110.

HARDING, D. J., MORENOFF, J. D., NGUYEN, A. P., BUSHWAY, S. D. and BINSWANGER, I. A. (2019). A natural experiment study of the effects of imprisonment on violence in the community. *Nat. Hum. Behav.* **3** 671–677. https://doi.org/10.1038/s41562-019-0604-8

HAUSMAN, C. H. J. and NEWEY, W. K. (2004). Many instruments, weak instruments, and microeonometric practice.

HERNÁN, M. A. and ROBINS, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology* **17** 360–372. https://doi.org/10.1097/01.ede.0000222409.00878.37

HICKERT, A., BUSHWAY, S. D., HARDING, D. J. and MORENOFF, J. D. (2022). Prior punishments and cumulative disadvantage: How supervision status impacts prison sentences. *Criminology* **60** 27–59.

HODGES, J. L. JR. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598–611. MR0152070 https://doi.org/10.1214/aoms/1177704172

IMBENS, G. W. (2014). Instrumental variables: An econometrician's perspective. *Statist. Sci.* **29** 323–358. MR3264545 https://doi.org/10.1214/14-STS480

INOUE, A. and SOLON, G. (2010). Two-sample instrumental variables estimators. *Rev. Econ. Stat.* **92** 557–561.

KADANE, J. B. and ANDERSON, T. W. (1977). A comment on the test of overidentifying restrictions. *Econometrica* **45** 1027–1031. MR0483260 https://doi.org/10.2307/1912691

KAFFO, M. and WANG, W. (2017). On bootstrap validity for specification testing with many weak instruments. *Econom. Lett.* **157** 107–111. MR3669092 https://doi.org/10.1016/j.econlet.2017.06.004

KANG, H., CAI, T. and SMALL, D. (2015). Robust confidence intervals for causal effects with possibly invalid instruments. Tech. rep., stat.wharton.upenn.

KLEIBERGEN, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica* **70** 1781–1803. MR1925156 https://doi.org/10.1111/1468-0262.00353

KLEIBERGEN, F. (2005). Testing parameters in GMM without assuming that they are identified. *Econometrica* **73** 1103–1123. MR2149242 https://doi.org/10.1111/j.1468-0262.2005.00610.x

KLEIBERGEN, F. (2007). Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices and identification statistics. *J. Econometrics* **139** 181–216. MR2380764 https://doi.org/10.1016/j.jeconom.2006.06.010

KLING, J. R. (2006). Incarceration length, employment, and earnings. Tech. Rep., National Bureau of Economic Research.

KOLESÁR, M., CHETTY, R., FRIEDMAN, J., GLAESER, E. and IMBENS, G. W. (2015). Identification and inference with many invalid instruments. *J. Bus. Econom. Statist.* **33** 474–484. MR3416595 https://doi.org/10.1080/07350015.2014.978175

LALONDE, R. J. and CHO, R. M. (2008). The impact of incarceration in state prison on the employment prospects of women. *J. Quant. Criminol.* **24** 243–265.

LANG, K. and KAHN-LANG SPITZER, A. (2020). Race discrimination: An economic perspective. *J. Econ. Perspect.* **34** 68–89.

LEVINE, B., MAHAR, A. and SMITH, J. (2021). Do Michigan's sentencing guidelines meet the legislature's goals?

LOEFFLER, C. E. (2013). Does imprisonment alter the life course? Evidence on crime and employment from a natural experiment. *Criminology* **51** 137–166.

MAGDALINOS, M. A. (1990). The classical principles of testing using instrumental variables estimates. *J. Econometrics* **44** 241–279. MR1060648 https://doi.org/10.1016/0304-4076(90)90059-3

MAGDALINOS, M. A. (1994). Testing instrument admissibility: Some refined asymptotic results. *Econometrica* **62** 373–403. MR1281696 https://doi.org/10.2307/2951617

MICHIGAN JUDICIAL INSTITUTE (2006). State of Michigan sentencing guidelines manual. Available at https://www.courts.michigan.gov/49c24d/siteassets/offices/mji/felony-sentencing-online-resources/2006_sgm.pdf.

MITCHELL, K. L. (2017). State sentencing guidelines: A garden full of variety. *Fed. Probat.* **81** 28.

MORGAN, S. L. and WINSHIP, C. (2007). *Counterfactuals and Causal Analysis*: *Methods and Principles for Social Research*.

NAGIN, D. S., CULLEN, F. T. and JONSON, C. L. (2009). Imprisonment and reoffending. *Crime Justice* **38** 115–200.

NELSON, C. and STARTZ, R. (1990a). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *J. Bus.* **63** S125–40.

NELSON, C. R. and STARTZ, R. (1990b). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica* **58** 967–976. MR1064853 https://doi.org/10.2307/2938359

NEWEY, W. K. and WINDMEIJER, F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica* **77** 687–719. MR2531359 https://doi.org/10.3982/ECTA6224

PAGER, D. (2008). *Marked*: *Race*, *Crime*, *and Finding Work in an Era of Mass Incarceration*. University Chicago Press, Chicago.

PETTIT, B. and LYONS, C. J. (2009). Incarceration and the legitimate labor market: Examining age-graded effects on employment and wages. *Law Soc*. *Rev.* **43** 725–756.

PUBLIC SAFETY PERFORMANCE PROJECT (2010). Prison Counts 2010: State Population Declines for the First Time in 38 Years. Washington, D.C.

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. *Springer Series in Statistics*. Springer, New York. MR2561612 https://doi.org/10.1007/978-1-4419-1213-8

SCHMITT, J., WARNER, K., GUPTA, S. et al. (2010). *The high budgetary cost of incarceration*. Center for Economic and Policy Research, Washington, DC.

SMALL, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J*. *Amer. Statist*. *Assoc*. **102** 1049–1058. MR2411664 https://doi.org/10.1198/016214507000000608

SMALL, D. S. and ROSENBAUM, P. R. (2008). War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *J*. *Amer. Statist*. *Assoc*. **103** 924–933. MR2528819 https://doi.org/10.1198/016214507000001247

STAIGER, D. and STOCK, J. H. (1997a). Instrumental variables regression with weak instruments. *Econometrica* **65** 557–586. MR1445622 https://doi.org/10.2307/2171753

STOCK, J. (2002). Instrumental variables in economics and statistics. *Int*. *Encyclopedia Soc*. *Sci*. 7577–7582.

STOCK, J. H. and WRIGHT, J. H. (2000). GMM with weak identification. *Econometrica* **68** 1055–1096. MR1779144 https://doi.org/10.1111/1468-0262.00151

STOCK, J. H., WRIGHT, J. H. and YOGO, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *J*. *Bus*. *Econom*. *Statist*. **20** 518–529. MR1973801 https://doi.org/10.1198/073500102288618658

STOCK, J. H. and YOGO, M. (2005). Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models* 80–108. Cambridge Univ. Press, Cambridge. MR2232140 https://doi.org/10.1017/CBO9780511614491.006

SWANSON, S. A., ROBINS, J. M., MILLER, M. and HERNÁN, M. A. (2015). Selecting on treatment: A pervasive form of bias in instrumental variable analyses. *Amer. J*. *Epidemiol*. **181** 191–197. https://doi.org/10.1093/aje/kwu284

WANG, J. and ZIVOT, E. (1998). Inference on structural parameters in instrumental variables regression with weak instruments. *Econometrica* **66** 1389–1404. MR1654303 https://doi.org/10.2307/2999621

WANG, X., JIANG, Y., ZHANG, N. R. and SMALL, D. S. (2018). Sensitivity analysis and power for instrumental variable studies. *Biometrics* **74** 1150–1160. MR3908133 https://doi.org/10.1111/biom.12873

WEIMAN, D. F., STOLL, M. A. and BUSHWAY, S. (2007). The regime of mass-incarceration: A labor-market perspective. In: *Barriers to Reentry?*: *The Labor Market for Released Prisoners in Post-Industrial America* (S. Bushway, M. A. Stoll and D. F. Weiman, eds.), 29–79 Russell Sage Foundation, New York.

WESTERN, B. (2002). The impact of incarceration on wage mobility and inequality. *Am*. *Sociol*. *Rev.* 526–546.

WESTERN, B., KLING, J. R. and WEIMAN, D. F. (2001). The labor market consequences of incarceration. *J*. *Res*. *Crime Delinq*. **47** 410–427.

WRIGHT, S. (1934). The method of path coefficients. *Ann*. *Math*. *Statist*. **5** 161–215.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 866–888
https://doi.org/10.1214/24-AOAS1994

ESTIMATING HETEROGENEOUS CAUSAL EFFECTS OF HIGH-DIMENSIONAL TREATMENTS: APPLICATION TO CONJOINT ANALYSIS

BY MAX GOPLERUD[1,a], KOSUKE IMAI[2,b] AND NICOLE E. PASHLEY[3,c]

[1]*Department of Government, University of Texas at Austin,* [a]*mgoplerud@austin.utexas.edu*
[2]*Department of Government and Department of Statistics, Harvard University,* [b]*imai@harvard.edu*
[3]*Department of Statistics, Rutgers University,* [c]*nicole.pashley@rutgers.edu*

Estimation of heterogeneous treatment effects is an active area of research. Most of the existing methods, however, focus on estimating the conditional average treatment effects of a single, binary treatment given a set of pretreatment covariates. In this paper we propose a method to estimate the heterogeneous causal effects of high-dimensional treatments, which poses unique challenges in terms of estimation and interpretation. The proposed approach finds maximally heterogeneous groups and uses a Bayesian mixture of regularized logistic regressions to identify groups of units who exhibit similar patterns of treatment effects. By directly modeling group membership with covariates, the proposed methodology allows one to explore the unit characteristics that are associated with different patterns of treatment effects. Our motivating application is conjoint analysis, which is a popular type of survey experiment in social science and marketing research and is based on a high-dimensional factorial design. We apply the proposed methodology to the conjoint data, where survey respondents are asked to select one of two immigrant profiles with randomly selected attributes. We find that a group of respondents with a relatively high degree of prejudice appears to discriminate against immigrants from non-European countries, like Iraq. An open-source software package is available for implementing the proposed methodology.

REFERENCES

ALMIRALL, D., GRIFFIN, B. A., MCCAFFREY, D. F., RAMCHAND, R., YUEN, R. A. and MURPHY, S. A. (2014). Time-varying effect moderation using the structural nested mean model: Estimation using inverse-weighted regression with residuals. *Stat. Med.* **33** 3466–3487.

ANDREWS, R. L., AINSLIE, A. and CURRIM, I. S. (2002). An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *J. Mark. Res.* **39** 479–487.

ATHEY, S. and IMBENS, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci.* **113** 7353–7360.

BONDELL, H. D. and REICH, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics* **65** 169–177. https://doi.org/10.1111/j.1541-0420.2008.01061.x

BRYANT, P. and WILLIAMSON, J. A. (1978). Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika* **65** 273–281.

CELEUX, G., FRÜHWIRTH-SCHNATTER, S. and ROBERT, C. P. (2019). Model selection for mixture models – perspectives and strategies. In *Handbook of Mixture Analysis* (S. Frühwirth-Schnatter, G. Celeux and C. P. Robert, eds.) 118–154. CRC Press/CRC, Boca Raton.

CHAMROUKHI, F. and HUYNH, B.-T. (2019). Regularized maximum likelihood estimation and feature selection in mixtures-of-experts models. *J. Soc. Fr. Stat.* **160** 57–85.

CHERNOZHUKOV, V., DEMIRER, M., DUFLO, E. and FERNANDEZ-VAL, I. (2023). Fisher-Schultz Lecture: Generic machine learning inference on heterogenous treatment effects in randomized experiments, with an application to immunization in India. Technical Report. Available at arXiv:1712.04802.

CHI, J. T., CHI, E. C. and BARANIUK, R. G. (2016). k-POD: A method for k-means clustering of missing data. *Amer. Statist.* **70** 91–99.

Key words and phrases. Causal inference, factorial design, mixture model, randomized experiment, regularized regression.

DASGUPTA, T., PILLAI, N. S. and RUBIN, D. B. (2015). Causal inference from 2^K factorial designs by using potential outcomes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 727–753.
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** 1–22.
DE LA CUESTA, B., EGAMI, N. and IMAI, K. (2022). Improving the external validity of conjoint analysis: The essential role of profile distribution. *Polit. Anal.* **30** 19–45.
EGAMI, N. and IMAI, K. (2019). Causal interaction in factorial experiments: Application to conjoint analysis. *J. Amer. Statist. Assoc.* **114** 529–540.
FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.
FIGUEIREDO, M. A. T. (2003). Adaptive sparseness for supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **25** 1150–1159.
GELMAN, A., HILL, J. and YAJIMA, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Eff.* **5** 189–211.
GOPLERUD, M. (2021). Modelling heterogeneity using Bayesian structured sparsity. Technical Report. Available at arXiv:2103.15919.
GOPLERUD, M., IMAI, K. and PASHLEY, N. E. (2025a). Replication data for "Estimating heterogeneous causal effects of high-dimensional treatments: Application to conjoint analysis." https://doi.org/10.7910/DVN/YAHPEH
GOPLERUD, M., IMAI, K. and PASHLEY, N. E (2025b). Supplement to "Estimating heterogeneous causal effects of high-dimensional treatments: Application to conjoint analysis." https://doi.org/10.1214/24-AOAS1994SUPPA, https://doi.org/10.1214/24-AOAS1994SUPPB
GOPLERUD, M., PASHLEY, N. E. and IMAI, K. (2025). FactorHet: Estimate heterogeneous effects in factorial experiments using grouping and sparsity. R package version 1.0.0. https://doi.org/10.32614/CRAN.package.FactorHet
GORMLEY, I. C. and FRÜHWIRTH-SCHNATTER, S. (2019). Mixture of experts models. In *Handbook of Mixture Analysis* (S. Frühwirth-Schnatter, G. Celeux and C. P. Robert, eds.) 271–307. Chapman and Hall/CRC.
GRIMMER, J., MESSING, S. and WESTWOOD, S. J. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Polit. Anal.* **25** 413–434.
GRÜN, B. and LEISCH, F. (2008). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *J. Classification* **25** 225–247.
GUPTA, S. and CHINTAGUNTA, P. K. (1994). On using demographic variables to determine segment membership in logit mixture models. *J. Mark. Res.* **31** 128–136.
HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Anal.* **15** 965–1056.
HAINMUELLER, J. and HOPKINS, D. J. (2015). The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants. *Amer. J. Polit. Sci.* **59** 529–548.
HAINMUELLER, J., HOPKINS, D. J. and YAMAMOTO, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Polit. Anal.* **22** 1–30.
IMAI, K. and LI, M. L. (2025). Statistical inference for heterogeneous treatment effects discovered by generic machine learning in randomized experiments. *J. Bus. Econom. Statist.* **43** 256–268.
IMAI, K. and RATKOVIC, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* **7** 443–470.
IMAI, K. and STRAUSS, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Polit. Anal.* **19** 1–19.
JIANG, W. and TANNER, M. A. (1999). On the identifiability of mixtures-of-experts. *Neural Netw.* **12** 1253–1258.
KHALILI, A. (2010). New estimation and feature selection methods in mixture-of-experts models. *Canad. J. Statist.* **38** 519–539.
KHALILI, A. and CHEN, J. (2007). Variable selection in finite mixture of regression models. *J. Amer. Statist. Assoc.* **102** 1025–1038.
KÜNZEL, S. R., SEKHON, J. S., BICKEL, P. J. and YU, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. USA* **116** 4156–4165.
LEEPER, T. J., HOBOLT, S. B. and TILLEY, J. (2020). Measuring subgroup preferences in conjoint experiments. *Polit. Anal.* **28** 207–221.
LIU, G. and SHIRAITO, Y. (2023). Multiple hypothesis testing in conjoint analysis. *Polit. Anal.* **31** 380–395.
LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44** 226–233.
MCLACHLAN, G. J. (1982). The classification and mixture maximum likelihood approaches to cluster analysis. In *Classification, Pattern Recognition and Reduction of Dimensionality*, (P. R. Krishnaiah and L. N. Kanal, eds.) **2** 199-208. North-Holland.

MENG, X.-L. and VAN DYK, D. A. (1997). The EM algorithm – an old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 511–567.

NEWMAN, B. J. and MALHOTRA, N. (2019). Economic reasoning with racial hue: Is the immigration consensus purely race neutral? *J. Polit.* **81** 153–166.

POLSON, N. G. and SCOTT, S. L. (2011). Data augmentation for support vector machines. *Bayesian Anal.* **6** 1–24.

POST, J. B. and BONDELL, H. D. (2013). Factor selection and structural identification in the interaction ANOVA model. *Biometrics* **69** 70–79.

RAO, V. R. (2014). *Applied Conjoint Analysis*. Springer, Berlin Heidelberg.

RATKOVIC, M. and TINGLEY, D. (2017). Sparse estimation and uncertainty with application to subgroup analysis. *Polit. Anal.* **25** 1–40.

ROBINSON, T. S. and DUCH, R. M. (2024). How to detect heterogeneity in conjoint experiments. *J. Polit.* **86** 412–427.

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

STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ-1-penalization for mixture regression models. *TEST* **19** 209–256.

STOKELL, B. G., SHAH, R. D. and TIBSHIRANI, R. J. (2021). Modelling high-dimensional categorical data using nonconvex fusion penalties. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 579–611.

TIAN, L., SLIZADEH, A. A., GENTLES, A. J. and TIBSHIRANI, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *J. Amer. Statist. Assoc.* **109** 1517–1532.

VAN DER LAAN, M. J. and ROSE, S. (2011b). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, Berlin.

WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242.

WU, C. J. and HAMADA, M. S. (2021). *Experiments: Planning, Analysis, and Optimization*, 3rd ed. Wiley, New York.

YAN, X. and BIEN, J. (2017). Hierarchical sparse modeling: A choice of two group lasso formulations. *Statist. Sci.* **32** 531–560.

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.

ZAHID, F. M. and TUTZ, G. (2013). Ridge estimation for multinomial logit models with symmetric side constraints. *Comput. Statist.* **28** 1017–1034.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 889–906
https://doi.org/10.1214/24-AOAS2005

EFFECTS OF ADOLESCENT VICTIMIZATION ON OFFENDING: FLEXIBLE METHODS FOR MISSING DATA AND UNMEASURED CONFOUNDING

BY MATEO DULCE RUBIO[1,2,a], EDWARD H. KENNEDY[1,b], VALERIO BAĆAK[3,d] AND DANIEL S. NAGIN[2,c]

[1]*Department of Statistics and Data Science, Carnegie Mellon University,* [a]*mdulceru@andrew.cmu.edu,* [b]*edward@stat.cmu.edu*

[2]*Heinz College of Information Systems and Public Policy, Carnegie Mellon University,* [c]*dn03@andrew.cmu.edu*

[3]*School of Criminal Justice, Rutgers University,* [d]*valerio.bacak@rutgers.edu*

The causal link between victimization and violence later in life is largely accepted but has been understudied for victimized adolescents. In this work we use the Add Health dataset, the largest nationally representative longitudinal survey of adolescents, to estimate the relationship between victimization and future offending in this population. To accomplish this, we derive a new doubly robust estimator for the average treatment effect on the treated (ATT) when the exposure and outcome are not always observed. We then find that the offending rate among victimized individuals would have been 3.86 percentage points lower if none of them had been victimized (95% CI: [0.28, 7.45]). This contributes positive evidence of a causal effect of victimization on future offending among adolescents. We further present statistical evidence of heterogeneous effects by age under which the ATT decreases according to the age at which victimization is experienced. We then devise a novel risk-ratio-based sensitivity analysis and conclude that our results are robust to modest unmeasured confounding. Finally, we show that the found effect is mainly driven by nonviolent offending.

REFERENCES

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189 https://doi.org/10.1111/j.1541-0420.2005.00377.x

BERG, M. T. and MULFORD, C. F. (2020). Reappraising and redirecting research on the victim–offender overlap. *Trauma Violence Abuse* **21** 16–30. https://doi.org/10.1177/1524838017735925

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

CHAUDHURI, S. and GUILKEY, D. K. (2016). GMM with multiple missing variables. *J*. *Appl*. *Econometrics* **31** 678–706. MR3512319 https://doi.org/10.1002/jae.2444

CHEN, P. and CHANTALA, K. (2014). Guidelines for analyzing Add Health data. Carolina Population Center, Univ. North Carolina at Chapel Hill 710.

CHEN, X. (2009). The link between juvenile offending and victimization: The influence of risky lifestyles, social bonding, and individual characteristics. *Youth Violence Juvenile Justice* **7** 119–135.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom*. *J*. **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CURRIE, J. and TEKIN, E. (2012). Understanding the cycle: Childhood maltreatment and future crime. *J*. *Hum*. *Resour*. **47** 509–549.

DAIGLE, L. E., BEAVER, K. M. and HARTMAN, J. L. (2008). A life-course approach to the study of victimization and offending behaviors. *Victim*. *Offend*. **3** 365–390.

DULCE RUBIO, M., KENNEDY, E. H., BAĆAK, V. and NAGIN, D. S. (2025). Supplement to "Effects of adolescent victimization on offending: flexible methods for missing data and unmeasured confounding." https://doi.org/10.1214/24-AOAS2005SUPP

HARRIS, K. M., HALPERN, C. T., WHITSEL, E. A., HUSSEY, J. M., KILLEYA-JONES, L. A., TABOR, J. and DEAN, S. C. (2019). Cohort profile: The national longitudinal study of adolescent to adult health (add health). *Int*. *J*. *Epidemiol*. **48** 1415–1415k. https://doi.org/10.1093/ije/dyz115

Key words and phrases. Victimization and offending, causal inference, doubly robust estimation, missing treatment.

HERNAN, M. A. and ROBINS, J. M. (2020). *Causal Inference: What If. Chapman & Hall/CRC Monographs on Statistics & Applied Probab.* CRC Press, Boca Raton.

HINES, O., DUKES, O., DIAZ-ORDAZ, K. and VANSTEELANDT, S. (2022). Demystifying statistical learning based on efficient influence functions. *Amer. Statist.* **76** 292–304. MR4453533 https://doi.org/10.1080/00031305.2021.2021984

IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.

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

KENNEDY, E. H. (2020). Efficient nonparametric causal inference with missing exposure information. *Int. J. Biostat.* **16**.

KENNEDY, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. arXiv preprint. Available at arXiv:2203.06469.

KENNEDY, E. H., KANGOVI, S. and MITRA, N. (2019). Estimating scaled treatment effects with multiple outcomes. *Stat. Methods Med. Res.* **28** 1094–1104. MR3934637 https://doi.org/10.1177/0962280217747130

KENNEDY, E. R. (2021). Package npcausal. Available at https://github.com/ehkennedy/npcausal.

LAURITSEN, J. L., SAMPSON, R. J. and LAUB, J. H. (1991). The link between offending and victimization among adolescents. *Criminology* **29** 265–292.

LITTLE, R. J. A. and RUBIN, D. B. (2019). *Statistical Analysis with Missing Data* 793. John Wiley & Sons.

LUEDTKE, A. R., DIAZ, I. and VAN DER LAAN, M. J. (2015). The statistics of sensitivity analyses.

MERSKY, J. P., TOPITZES, J. and REYNOLDS, A. J. (2012). Unsafe at any age: Linking childhood and adolescent maltreatment to delinquency and crime. *J. Res. Crime Delinq.* **49** 295–318. https://doi.org/10.1177/0022427811415284

MYERS, W., LLOYD, K., TURANOVIC, J. J. and PRATT, T. C. (2018). Revisiting a criminological classic: The cycle of violence. *J. Contemp. Crim. Justice* **34** 266–286.

OJJDP (2013). Bullying. Literature Review. [Online].

OJJDP (2018). OJJDP Statistical Briefing Book.

OLWEUS, D. (2013). School bullying: Development and some important challenges. *Annu. Rev. Clin. Psychol.* **9** 751–780. https://doi.org/10.1146/annurev-clinpsy-050212-185516

TENEYCK, M. and BARNES, J. (2018). Exploring the social and individual differences among victims, offenders, victim–offenders, and total abstainers. *Victim. Offend.* **13** 66–83.

TTOFI, M. M., FARRINGTON, D. P. and LÖSEL, F. (2012). School bullying as a predictor of violence later in life: A systematic review and meta-analysis of prospective longitudinal studies. *Aggress. Violent Behav.* **17** 405–418.

VANDERWEELE, T. J. and HERNAN, M. A. (2013). Causal inference under multiple versions of treatment. *J. Causal Inference* **1** 1–20. MR4289399 https://doi.org/10.1515/jci-2012-0002

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** Art. 25, 23. MR2349918 https://doi.org/10.2202/1544-6115.1309

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality. Springer Series in Statistics*. Springer, New York. MR1958123 https://doi.org/10.1007/978-0-387-21700-0

VAN DER VAART, A. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1999). Lecture Notes in Math.* **1781** 331–457. Springer, Berlin. MR1915446

WATTS, S. J. (2017). The link between child abuse and neglect and delinquency: Examining the mediating role of social bonds. *Victim. Offend.* **12** 700–717.

WIDOM, C. S. (1989). The cycle of violence. *Science* **244** 160–166. https://doi.org/10.1126/science.2704995

WIDOM, C. S., CZAJA, S. J. and DUMONT, K. A. (2015). Intergenerational transmission of child abuse and neglect: Real or detection bias? *Science* **347** 1480–1485. https://doi.org/10.1126/science.1259917

WIDOM, C. S. and MAXFIELD, M. G. (2001). An Update on the 'Cycle of Violence'. Research in Brief.

WILLIAMSON, E. J., FORBES, A. and WOLFE, R. (2012). Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder. *Stat. Med.* **31** 4382–4400. MR3040088 https://doi.org/10.1002/sim.5643

WRIGHT, E. M. and FAGAN, A. A. (2013). The cycle of violence in context: Exploring the moderating roles of neighborhood disadvantage and cultural norms. *Criminology* **51** 217–249. https://doi.org/10.1111/1745-9125.12003

ZHANG, Z., LIU, W., ZHANG, B., TANG, L. and ZHANG, J. (2016). Causal inference with missing exposure information: Methods and applications to an obstetric study. *Stat. Methods Med. Res.* **25** 2053–2066. MR3553326 https://doi.org/10.1177/0962280213513758

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 907–925
https://doi.org/10.1214/25-AOAS2023

THE PROPERTIES OF COVARIATE-ADAPTIVE RANDOMIZATION PROCEDURES WITH POSSIBLY UNEQUAL ALLOCATION RATIO

BY XIAO LIU[1,a], FEIFANG HU[2,c] AND WEI MA[1,b]

[1]*Institute of Statistics and Big Data, Renmin University of China,* [a]*lx0786@ruc.edu.cn,* [b]*mawei@ruc.edu.cn*
[2]*Department of Statistics, George Washington University,* [c]*feifang@gwu.edu*

In clinical trials, covariate-adaptive randomization (CAR) procedures are used to balance important covariates for more convincing results and enhanced statistical efficiency. Although most CAR procedures focus on trials with 1:1 allocation ratio, the demand for unequal allocation is growing. Therefore, this paper proposes a CAR framework that unifies numerous existing procedures and can balance general (discrete, continuous, or their combinations) covariates under any allocation ratio. To evaluate the proposed procedure, we classify covariates into randomized covariates and additional covariates, based on whether or not they are used in the randomization procedure. The analysis indicates that our proposed procedure possesses superior balancing properties for randomized covariates. Subsequently, we investigate the impact of CAR procedures on additional covariates. When balancing only discrete covariates, our results exhibit the benefit of CAR procedures in balancing additional covariates. However, the most intriguing finding is that, under unequal allocation ratio, balancing continuous covariates will challenge the balance of additional covariates, and we refer to this new issue as the "shift problem." To understand and remedy this issue, we perform a comprehensive analysis about when and why it occurs, followed by two practical solutions to address the shift problem. The proposed CAR procedures are shown to effectively balance covariates when applied to the data from a depression trial.

REFERENCES

AICKIN, M. (2001). Randomization, balance, and the validity and efficiency of design-adaptive allocation methods. *J. Statist. Plann. Inference* **94** 97–119. MR1820173 https://doi.org/10.1016/S0378-3758(00)00228-7

BRUHN, M. and MCKENZIE, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *Am. Econ. J. Appl. Econ.* **1** 200–32.

DUMVILLE, J. C., HAHN, S., MILES, J. N. V. and TORGERSON, D. J. (2006). The use of unequal randomisation ratios in clinical trials: A review. *Contemp. Clin. Trials* **27** 1–12. https://doi.org/10.1016/j.cct.2005.08.003

EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58** 403–417. MR0312660 https://doi.org/10.1093/biomet/58.3.403

ENDO, A., NAGATANI, F., HAMADA, C. and YOSHIMURA, I. (2006). Minimization method for balancing continuous prognostic variables between treatment and control groups using Kullback-Leibler divergence. *Contemp. Clin. Trials* **27** 420–431. https://doi.org/10.1016/j.cct.2006.05.002

HAN, B., ENAS, N. H. and MCENTEGART, D. (2009). Randomization by minimization for unbalanced treatment allocation. *Stat. Med.* **28** 3329–3346. MR2744367 https://doi.org/10.1002/sim.3710

HAN, B., YU, M. and MCENTEGART, D. (2013). Weighted re-randomization tests for minimization with unbalanced allocation. *Pharm. Stat.* **12** 243–53.

HOEHLER, F. K. (1987). Balancing allocation of subjects in biomedical research: A minimization strategy based on ranks. *Comput. Biomed. Res.* **20** 209–213. https://doi.org/10.1016/0010-4809(87)90054-1

HU, F., YE, X. and ZHANG, L.-X. (2023). Multi-arm covariate-adaptive randomization. *Sci. China Math.* **66** 163–190. MR4529032 https://doi.org/10.1007/s11425-020-1954-y

HU, Y. and HU, F. (2012). Asymptotic properties of covariate-adaptive randomization. *Ann. Statist.* **40** 1794–1815. MR3015044 https://doi.org/10.1214/12-AOS983

Key words and phrases. Covariate-adaptive randomization, unequal allocation ratio, covariate balance, balancing additional covariates, shift problem.

KAISER, L. D. (2012). Dynamic randomization and a randomization model for clinical trials data. *Stat. Med.* **31** 3858–3873. MR3041778 https://doi.org/10.1002/sim.5448

KELLER, M. B., MCCULLOUGH, J. P., KLEIN, D. N., ARNOW, B., DUNNER, D. L., GELENBERG, A. J., MARKOWITZ, J. C., NEMEROFF, C. B., RUSSELL, J. M. et al. (2000). A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. *N. Engl. J. Med.* **342** 1462–70.

KUNDT, G. (2009). Comparative evaluation of balancing properties of stratified randomization procedures. *Methods Inf. Med.* **48** 129–134. https://doi.org/10.3414/ME0538

KUZNETSOVA, O. M. and JOHNSON, V. P. (2017). Approaches to expanding the two-arm biased coin randomization to unequal allocation while preserving the unconditional allocation ratio. *Stat. Med.* **36** 2483–2498. MR3660145 https://doi.org/10.1002/sim.7290

KUZNETSOVA, O. M. and TYMOFYEYEV, Y. (2012). Preserving the allocation ratio at every allocation with biased coin randomization and minimization in studies with unequal allocation. *Stat. Med.* **31** 701–723. MR2901795 https://doi.org/10.1002/sim.4447

LI, X., ZHOU, J. and HU, F. (2019). Testing hypotheses under adaptive randomization with continuous covariates in clinical trials. *Stat. Methods Med. Res.* **28** 1609–1621. MR3961953 https://doi.org/10.1177/0962280218770231

LIU, X., HU, F. and MA, W. (2025). Supplement to "The properties of covariate-adaptive randomization procedures with possibly unequal allocation ratio." https://doi.org/10.1214/25-AOAS2023SUPPA, https://doi.org/10.1214/25-AOAS2023SUPPB

LIU, Y. and HU, F. (2022). Balancing unobserved covariates with covariate-adaptive randomized experiments. *J. Amer. Statist. Assoc.* **117** 875–886. MR4436319 https://doi.org/10.1080/01621459.2020.1825450

LIU, Z., BAN, T. and HUANG, T. (2017). General covariate-adaptive randomization targeting unequal allocation ratio. *J. Statist. Plann. Inference* **191** 68–80. MR3679109 https://doi.org/10.1016/j.jspi.2017.05.008

MA, W., LI, P., ZHANG, L.-X. and HU, F. (2024). A new and unified family of covariate adaptive randomization procedures and their properties. *J. Amer. Statist. Assoc.* **119** 151–162. MR4713882 https://doi.org/10.1080/01621459.2022.2102986

MA, Z. and HU, F. (2013). Balancing continuous covariates based on kernel densities. *Contemp. Clin. Trials* **34** 262–9.

MCENTEGART, D. J. (2003). The pursuit of balance using stratified and dynamic randomization techniques: An overview. *Drug Inf. J.* **37** 293–308.

POCOCK, S. J. and SIMON, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **31** 103–115.

PROSCHAN, M., BRITTAIN, E. and KAMMERMAN, L. (2011). Minimize the use of minimization with unequal allocation. *Biometrics* **67** 1135–1141. MR2829249 https://doi.org/10.1111/j.1541-0420.2010.01545.x

QIN, Y., LI, Y., MA, W., YANG, H. and HU, F. (2024). Adaptive randomization via Mahalanobis distance. *Statist. Sinica* **34** 353–375. MR4683576 https://doi.org/10.5705/ss.202020.0440

ROSENBERGER, W. F. and LACHIN, J. M. (2016). *Randomization in Clinical Trials: Theory and Practice*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3443072 https://doi.org/10.1002/9781118742112

SHAO, J., YU, X. and ZHONG, B. (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika* **97** 347–360. MR2650743 https://doi.org/10.1093/biomet/asq014

STIGSBY, B. and TAVES, D. R. (2010). Rank-minimization for balanced assignment of subjects in clinical trials. *Contemp. Clin. Trials* **31** 147–150. https://doi.org/10.1016/j.cct.2009.12.001

SU, Z. (2011). Balancing multiple baseline characteristics in randomized clinical trials. *Contemp. Clin. Trials* **32** 547–550. https://doi.org/10.1016/j.cct.2011.03.004

TAVES, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clin. Pharmacol. Ther.* **15** 443–53.

WANG, T. and MA, W. (2021). The impact of misclassification on covariate-adaptive randomized clinical trials. *Biometrics* **77** 451–464. MR4307647 https://doi.org/10.1111/biom.13308

WANG, T. and MA, W. (2025). The impact of misclassification on covariate-adaptive randomized clinical trials with generalized linear models. *J. Statist. Plann. Inference*. **234**. 106209. https://doi.org/10.1016/j.jspi.2024.106209.

ZELEN, M. (1974). The randomization and stratification of patients to clinical trials. *J. Chronic Dis.* **27** 365–75.

ZHANG, L.-X. (2023). Asymptotic properties of multi-treatment covariate adaptive randomization procedures for balancing observed and unobserved covariates. Available at arXiv:2305.13842.

ZHAO, Z., SONG, Y., JIANG, W. and TU, D. (2024). Consistent covariances estimation for stratum imbalances under minimization method for covariate-adaptive randomization. *Scand. J. Stat.* **51** 861–890. MR4764758 https://doi.org/10.1111/sjos.12703

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 926–942
https://doi.org/10.1214/24-AOAS1973

A REWEIGHTED RANDOM FOREST TO PREDICT HEALTH OUTCOMES USING HUMAN MICROBIOME DATA

BY TIAN WANG[1,a], BING LI[2,f], HUANG XU[1,b], YUQI MIAO[1,c], MIN QIAN[1,d] AND SHUANG WANG[1,e]

[1]*Department of Biostatistics, Mailman School of Public Health, Columbia University,* [a]*tw2697@cumc.columbia.edu,* [b]*hx2397@cumc.columbia.edu,* [c]*ym2771@cumc.columbia.edu,* [d]*mq2158@cumc.columbia.edu,* [e]*sw2206@columbia.edu*

[2]*Department of Biostatistics, Brown University,* [f]*bing_li1@alumni.brown.edu*

Many statistical methods examining associations and predictions of microbiome on health outcomes are distance-based, where several distance metrics are calculated to capture different aspects of microbiome and the "optimal" one is selected for final association or prediction. Studies have suggested that diverse forms of taxa are linked to health outcomes; that is, both abundant taxa in close proximity and rare taxa far away on the phylogenetic tree could be associated with or predictive of the same outcome. However, existing prediction methods often utilize only one association form. Among popular prediction models, random forest (RF) has been widely applied and shown satisfactory performance. Here we introduce the reweighted-RF estimate that reweights sample contributions based on their microbiome similarities, and develop multiple reweighted-RF estimates by adopting different distance metrics to encompass various microbiome-outcome association forms. These reweighted-RF estimates are then ensembled for final predictions using multiple signal types. In simulation studies we demonstrated improved prediction performance of the reweighted-RF estimate, with weights from the distance metric capturing the true microbiome signal form, over that of the original RF. The ensemble estimate also consistently outperforms the original RF or performs as well as the best reweighted-RF estimate when there exist multiple signal forms, where ensemble weights provide insights into the contributing signal forms. We applied our method to predict the binary obesity and irritable bowel syndrome and continuous body mass index and age, using both gut and oral microbiome data from the American Gut Project, and observed improved prediction results over those of competing methods.

REFERENCES

BIAU, G. and SCORNET, E. (2016). A random forest guided tour. *TEST* **25** 197–227. MR3493512 https://doi.org/10.1007/s11749-016-0481-7

BRAY, J. R. and CURTIS, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* **27** 325–349.

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

CHARLSON, E. S., CHEN, J., CUSTERS-ALLEN, R., BITTINGER, K., LI, H., SINHA, R., HWANG, J., BUSHMAN, F. D. and COLLMAN, R. G. (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS ONE* **5** e15216.

CHEN, J., BITTINGER, K., CHARLSON, E. S., HOFFMANN, C., LEWIS, J., WU, G. D., COLLMAN, R. G., BUSHMAN, F. D. and LI, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* **28** 2106–2113.

CHEN, L., LIU, H., KOCHER, J.-P. A., LI, H. and CHEN, J. (2015). glmgraph: An R package for variable selection and predictive modeling of structured genomic data. *Bioinformatics* **31** 3991–3993.

COLLADO, M. C., ISOLAURI, E., LAITINEN, K. and SALMINEN, S. (2008). Distinct composition of gut microbiota during pregnancy in overweight and normal-weight women. *Am. J. Clin. Nutr.* **88** 894–899. https://doi.org/10.1093/ajcn/88.4.894

Key words and phrases. Reweighted-random forest, microbiome, distance metrics.

FRIEDBERG, R., TIBSHIRANI, J., ATHEY, S. and WAGER, S. (2021). Local linear forests. *J. Comput. Graph. Statist.* **30** 503–517. MR4270520 https://doi.org/10.1080/10618600.2020.1831930

FUKUYAMA, J. (2019). Emphasis on the deep or shallow parts of the tree provides a new characterization of phylogenetic distances. *Genome Biol.* **20** 1–16.

GILIBERTI, R., CAVALIERE, S., MAURIELLO, I. E., ERCOLINI, D. and PASOLLI, E. (2022). Host phenotype classification from human microbiome data is mainly driven by the presence of microbial taxa. *PLoS Comput. Biol.* **18** e1010066.

KARLSSON, F. H., TREMAROLI, V., NOOKAEW, I., BERGSTRÖM, G., BEHRE, C. J., FAGERBERG, B., NIELSEN, J. and BÄCKHED, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498** 99–103. https://doi.org/10.1038/nature12198

KIM, K. J., PARK, J., PARK, S.-C. and WON, S. (2020). Phylogenetic tree-based microbiome association test. *Bioinformatics* **36** 1000–1006.

KNIGHTS, D., COSTELLO, E. K. and KNIGHT, R. (2011). Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35** 343–359. https://doi.org/10.1111/j.1574-6976.2010.00251.x

KOH, H., BLASER, M. J. and LI, H. (2017). A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. *Microbiome* **5** 45.

KOSLOVSKY, M. D., HOFFMAN, K. L., DANIEL, C. R. and VANNUCCI, M. (2020). A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. *Ann. Appl. Stat.* **14** 1471–1492. MR4152142 https://doi.org/10.1214/20-AOAS1354

LASKEN, R. S. (2012). Genomic sequencing of uncultured microorganisms from single cells. *Nat. Rev., Microbiol.* **10** 631–640. https://doi.org/10.1038/nrmicro2857

LIN, Y. and JEON, Y. (2006). Random forests and adaptive nearest neighbors. *J. Amer. Statist. Assoc.* **101** 578–590. MR2256176 https://doi.org/10.1198/016214505000001230

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

LOZUPONE, C. A., HAMADY, M., KELLEY, S. T. and KNIGHT, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.* **73** 1576–1585. https://doi.org/10.1128/AEM.01996-06

LU, J., SHI, P. and LI, H. (2019). Generalized linear models with linear constraints for microbiome compositional data. *Biometrics* **75** 235–244. MR3953724 https://doi.org/10.1111/biom.12956

MA, Y., LUO, Y. and JIANG, H. (2020). A novel normalization and differential abundance test framework for microbiome data. *Bioinformatics* **36** 3959–3965.

MCDONALD, D., HYDE, E., DEBELIUS, J. W., MORTON, J. T., GONZALEZ, A., ACKERMANN, G., AKSENOV, A. A., BEHSAZ, B., BRENNAN, C. et al. (2018). American gut: An open platform for citizen science microbiome research. *mSystems* **3**.

MCDONALD, D., PRICE, M. N., GOODRICH, J., NAWROCKI, E. P., DESANTIS, T. Z., PROBST, A., ANDERSEN, G. L., KNIGHT, R. and HUGENHOLTZ, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6** 610–618. https://doi.org/10.1038/ismej.2011.139

MICCHELLI, C. A., XU, Y. and ZHANG, H. (2006). Universal kernels. *J. Mach. Learn. Res.* **7** 2651–2667. MR2274454

MORGAN, X. C., TICKLE, T. L., SOKOL, H., GEVERS, D., DEVANEY, K. L., WARD, D. V., REYES, J. A., SHAH, S. A., LELEIKO, N. et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13** R79. https://doi.org/10.1186/gb-2012-13-9-r79

MUSCOGIURI, G., CANTONE, E., CASSARANO, S., TUCCINARDI, D., BARREA, L., SAVASTANO, S. and COLAO, A. (2019). Gut microbiota: A new path to treat obesity. *Int. J. Obes. Suppl.* **9** 10–19.

PRICE, M. N., DEHAL, P. S. and ARKIN, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5** e9490. https://doi.org/10.1371/journal.pone.0009490

RAPIN, A., PATTARONI, C., MARSLAND, B. J. and HARRIS, N. L. (2017). Microbiota analysis using an illumina MiSeq platform to sequence 16S rRNA genes. *Curr. Protoc. Mouse Biol.* **7** 100–129. https://doi.org/10.1002/cpmo.29

ROLF, E., JORDAN, M. I. and RECHT, B. (2020). Post-estimation smoothing: A simple baseline for learning with side information. In *International Conference on Artificial Intelligence and Statistics* 1759–1769. PMLR.

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

SONG, Y., ZHAO, H. and WANG, T. (2020). An adaptive independence test for microbiome community data. *Biometrics* **76** 414–426. MR4125267 https://doi.org/10.1111/biom.13154

SUBRAMANIAN, S., HUQ, S., YATSUNENKO, T., HAQUE, R., MAHFUZ, M., ALAM, M. A., BENEZRA, A., DESTEFANO, J., MEIER, M. F. et al. (2014). Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* **510** 417.

TANASEICHUK, O., BORNEMAN, J. and JIANG, T. (2014). Phylogeny-based classification of microbial communities. *Bioinformatics* **30** 449–456.

TANG, Y., MA, L. and NICOLAE, D. L. (2018). A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data. *Ann. Appl. Stat.* **12** 1–26. MR3773384 https://doi.org/10.1214/17-AOAS1086

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

TRINGE, S. G. and RUBIN, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* **6** 805–814. https://doi.org/10.1038/nrg1709

TURNBAUGH, P. J., HAMADY, M., YATSUNENKO, T., CANTAREL, B. L., DUNCAN, A., LEY, R. E., SOGIN, M. L., JONES, W. J., ROE, B. A. et al. (2009). A core gut microbiome in obese and lean twins. *Nature* **457** 480–484.

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** 25. MR2349918 https://doi.org/10.2202/1544-6115.1309

VAN HUL, M. and CANI, P. D. (2023). The gut microbiota in obesity and weight management: Microbes as friends or foe? *Nat. Rev. Endocrinol.* 1–14.

WANG, T., LI, B., XU, H., MIAO, Y., QIAN, M. and WANG, S. (2025). Supplement to "A reweighted random forest to predict health outcomes using human microbiome data." https://doi.org/10.1214/24-AOAS1973SUPPA, https://doi.org/10.1214/24-AOAS1973SUPPB

WANG, T., LING, W., PLANTINGA, A. M., WU, M. C. and ZHAN, X. (2021). Testing microbiome association using integrated quantile regression models. *Bioinformatics*.

WANG, T. and ZHAO, H. (2017a). Constructing predictive microbial signatures at multiple taxonomic levels. *J. Amer. Statist. Assoc.* **112** 1022–1031. MR3735357 https://doi.org/10.1080/01621459.2016.1270213

WANG, T. and ZHAO, H. (2017b). Structured subcomposition selection in regression and its application to microbiome data analysis. *Ann. Appl. Stat.* **11** 771–791. MR3693546 https://doi.org/10.1214/16-AOAS1017

WASSAN, J. T., WANG, H., BROWNE, F. and ZHENG, H. (2018). PAAM-ML: A novel phylogeny and abundance aware machine learning modelling approach for microbiome classification. In 2018 *IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*) 44–49. IEEE, Los Alamitos.

WU, C., CHEN, J., KIM, J. and PAN, W. (2016). An adaptive association test for microbiome data. *Gen. Med.* **8** 56.

XIAO, J., CHEN, L., JOHNSON, S., YU, Y., ZHANG, X. and CHEN, J. (2018). Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. *Front. Microbiol.* **9** 1391.

ZHAN, X., XUE, L., ZHENG, H., PLANTINGA, A., WU, M. C., SCHAID, D. J., ZHAO, N. and CHEN, J. (2018). A small-sample kernel association test for correlated data with application to microbiome association studies. *Genet. Epidemiol.* **42** 772–782.

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

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 943–966
https://doi.org/10.1214/24-AOAS1995

BIOMARKER DETECTION FOR DISEASE CLASSIFICATION IN LONGITUDINAL MICROBIOME DATA

BY CHAO CHENG[1], HANTENG MA[1], YUJIE ZHONG[2,a], ANNE-CATRIN UHLEMANN[3], XINGDONG FENG[1] AND JIANHUA HU[4,b]

[1]*School of Statistics and Data Science, Shanghai University of Finance and Economics*

[2]*AstraZeneca R&D China*, [a]*sabrina_zyj@163.com*

[3]*Division of Infectious Diseases, Department of Medicine, Columbia University*

[4]*Department of Biostatistics, Columbia University*, [b]*jh3992@cumc.columbia.edu*

The microbiome has been found to have a close relationship with human health. Advancements in sequencing technologies have enabled in-depth studies of microbial communities and their associations with various diseases. When analyzing microbiome data, it is common to perform compositional scale normalization to ensure statistical validity. This requires special treatment to address the unique characteristics of microbiome data. Furthermore, biomedical studies often involve repeated measurements of microbial samples, which adds complexity to the data analysis. In this paper we focus on a liver transplant microbiome study. The main objective is to investigate the association between the colonization status of multidrug-resistant bacteria (MDRB) and the longitudinal microbial abundance profile. To accomplish this, we employ a regularized functional logistic regression model in our analysis. Specifically, we utilize the log-contrast model with a low-rank approximation to handle the compositional covariates and nonconvex penalties to select the important components in the covariate space. We propose an efficient estimation algorithm and establish the oracle property of the estimator. We name this new development as Functional Compositional data Quadratic Method (FCQM). We demonstrate the promise of the proposed method with extensive simulation studies and the liver transplant application.

REFERENCES

AFRIDI, O. K., ALI, J. and CHANG, J. H. (2021). Resistome and microbial profiling of pediatric patient's gut infected with multidrug-resistant diarrhoeagenic Enterobacteriaceae using next-generation sequencing; the first study from Pakistan. *Libyan J. Med.* **16**.

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

AITCHISON, J. and BACON-SHONE, J. (1984). Log contrast models for experiments with mixtures. *Biometrika* **71** 323–330.

AITCHISON, J., BARCELÓ-VIDAL, C., MARTÍN-FERNÁNDEZ, J. A. and PAWLOWSKY-GLAHN, V. (2000). Logratio analysis and compositional distance. *Math. Geol.* **32** 271–275.

ANCONA, G., ALAGNA, L., LOMBARDI, A., PALOMBA, E., CASTELLI, V., RENISI, G., DONDOSSOLA, D., IAVARONE, M., MUSCATELLO, A. et al. (2021). The interplay between gut microbiota and the immune system in liver transplant recipients and its role in infections. *Infect. Immun.* **89** e0037621. https://doi.org/10.1128/IAI.00376-21

ANEIROS, G., FERRATY, F. and VIEU, P. (2015). Variable selection in partial linear regression with functional covariate. *Statistics* **49** 1322–1347. MR3415222 https://doi.org/10.1080/02331888.2014.998675

ANNAVAJHALA, M. K., GOMEZ-SIMMONDS, A., MACESIC, N., SULLIVAN, S. B., KRESS, A., KHAN, S. D., GIDDINS, M. J., STUMP, S., KIM, G. I. et al. (2019). Colonizing multidrug-resistant bacteria and the longitudinal evolution of the intestinal microbiome after liver transplantation. *Nat. Commun.* **10**.

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

Key words and phrases. Compositional data, functional data analysis, high-dimensional data, logistic regression.

BREHENY, P. J. (2019). Marginal false discovery rates for penalized regression models. *Biostatistics* **20** 299–314. MR3922135 https://doi.org/10.1093/biostatistics/kxy004

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

CHEN, C., LIAO, J., XIA, Y., LIU, X., JONES, R., HARAN, J., MCCORMICK, B., SAMPSON, T. R., ALAM, A. et al. (2022). Gut microbiota regulate Alzheimer's disease pathologies and cognitive disorders via PUFA-associated neuroinflammation. *Gut* gutjnl–2021–326269.

CHENG, C., FENG, X., LI, X. and WU, M. (2022). Robust analysis of cancer heterogeneity for high-dimensional data. *Stat. Med.* **41** 5448–5462. MR4511079 https://doi.org/10.1002/sim.9578

CHENG, C., MA, H., ZHONG, Y., UHLEMANN, A.-C., FENG, X. and HU, J. (2025). Supplement to "Biomarker detection for disease classification in longitudinal microbiome data." https://doi.org/10.1214/24-AOAS1995SUPPA, https://doi.org/10.1214/24-AOAS1995SUPPB

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194 https://doi.org/10.1214/009053604000000256

FAN, Y., JAMES, G. M. and RADCHENKO, P. (2015). Functional additive regression. *Ann. Statist.* **43** 2296–2325. MR3396986 https://doi.org/10.1214/15-AOS1346

FERREIRA, R. B. R., GILL, N., WILLING, B. P., ANTUNES, L. C. M., RUSSELL, S. L., CROXEN, M. A. and FINLAY, B. B. (2011). The intestinal microbiota plays a role in salmonella-induced colitis independent of pathogen colonization. *PLoS ONE* **6** e20338.

GLOOR, G. B., MACKLAIM, J. M., PAWLOWSKY-GLAHN, V. and EGOZCUE, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8**.

GLOOR, G. B. and REID, G. (2016). Compositional analysis: A valid approach to analyze microbiome high-throughput sequencing data. *Can. J. Microbiol.* **62** 692–703. https://doi.org/10.1139/cjm-2015-0821

HUANG, L., ZHAO, J., WANG, H. and WANG, S. (2016). Robust shrinkage estimation and selection for functional multiple linear model through LAD loss. *Comput. Statist. Data Anal.* **103** 384–400. MR3522639 https://doi.org/10.1016/j.csda.2016.05.017

HUNTER, D. R. and LANGE, K. (2000). Quantile regression via an MM algorithm. *J. Comput. Graph. Statist.* **9** 60–77. MR1819866 https://doi.org/10.2307/1390613

HUNTER, D. R. and LANGE, K. (2004). A tutorial on MM algorithms. *Amer. Statist.* **58** 30–37. MR2055509 https://doi.org/10.1198/0003130042836

JAMES, G. M., WANG, J. and ZHU, J. (2009). Functional linear regression that's interpretable. *Ann. Statist.* **37** 2083–2108. MR2543686 https://doi.org/10.1214/08-AOS641

KONSTANTINIDIS, K. T., ROSSELLÓ-MÓRA, R. and AMANN, R. (2017). Uncultivated microbes in need of their own taxonomy. *ISME J.* **11** 2399–2406. https://doi.org/10.1038/ismej.2017.113

KRISS, M., VERNA, E. C., ROSEN, H. R. and LOZUPONE, C. A. (2019). Functional microbiomics in liver transplantation: Identifying novel targets for improving allograft outcomes. *Transplantation* **103** 668–678. https://doi.org/10.1097/TP.0000000000002568

LANGE, K., HUNTER, D. R. and YANG, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graph. Statist.* **9** 1–59. MR1819865 https://doi.org/10.2307/1390605

LI, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.* **2** 73–94.

LIAN, H. (2013). Shrinkage estimation and selection for multiple functional regression. *Statist. Sinica* **23** 51–74. MR3076158

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

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

LU, J., SHI, P. and LI, H. (2019). Generalized linear models with linear constraints for microbiome compositional data. *Biometrics* **75** 235–244. MR3953724 https://doi.org/10.1111/biom.12956

MA, S. and HUANG, J. (2017). A concave pairwise fusion approach to subgroup analysis. *J. Amer. Statist. Assoc.* **112** 410–423. MR3646581 https://doi.org/10.1080/01621459.2016.1148039

MACESIC, N., NELSON, B., MCCONVILLE, T. H., GIDDINS, M. J., GREEN, D. A., STUMP, S., GOMEZ-SIMMONDS, A., ANNAVAJHALA, M. K. and UHLEMANN, A.-C. (2019). Emergence of polymyxin resistance in clinical klebsiella pneumoniae through diverse genetic adaptations: A genomic, retrospective cohort study. *Clin. Infect. Dis.* **70** 2084–2091.

MATSEN, F. A., KODNER, R. B. and ARMBRUST, E. V. (2010). pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinform*. **11** 538. https://doi.org/10.1186/1471-2105-11-538

MATSUI, H. and KONISHI, S. (2011). Variable selection for functional regression models via the L_1 regularization. *Comput*. *Statist*. *Data Anal*. **55** 3304–3310. MR2825412 https://doi.org/10.1016/j.csda.2011.06.016

NEARING, J. T., DOUGLAS, G. M., HAYES, M. G., MACDONALD, J., DESAI, D. K., ALLWARD, N., JONES, C. M. A., WRIGHT, R. J., DHANANI, A. S. et al. (2022). Microbiome differential abundance methods produce different results across 38 datasets. *Nat*. *Commun*. **13**.

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

QUINN, T. P., ERB, I., RICHARDSON, M. F. and CROWLEY, T. M. (2018). Understanding sequencing data as compositions: An outlook and review. *Bioinformatics* **34** 2870–2878. https://doi.org/10.1093/bioinformatics/bty175

RAMSAY, J. O. and SILVERMAN, B. W. (2002). *Applied Functional Data Analysis*: *Methods and Case Studies*. *Springer Series in Statistics*. Springer, New York. MR1910407 https://doi.org/10.1007/b98886

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

SCHUBERT, A. M., SINANI, H. and SCHLOSS, P. D. (2015). Antibiotic-induced alterations of the murine gut microbiota and subsequent effects on colonization resistance against clostridium difficile. *mBio* **6** e00974. https://doi.org/10.1128/mBio.00974-15

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

SIMON, N. and TIBSHIRANI, R. (2012). Standardization and the group Lasso penalty. *Statist*. *Sinica* **22** 983–1001. MR2987480 https://doi.org/10.5705/ss.2011.075

SIVARAJ, S., CHAN, A., PASINI, E., CHEN, E., LAWENDY, B., VERNA, E., WATT, K. and BHAT, M. (2020). Enteric dysbiosis in liver and kidney transplant recipients: A systematic review. *Transp*. *Int*. **33** 1163–1176.

STOREY, J. D., XIAO, W., LEEK, J. T., TOMPKINS, R. G. and DAVIS, R. W. (2005). Significance analysis of time course microarray experiments. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **102** 12837–12842.

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

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

TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., BROWN, P., HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D. and ALTMAN, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* **17** 520–525.

WANG, H., WANG, Q., YANG, C., GUO, M., CUI, X., JING, Z., LIU, Y., QIAO, W., QI, H., ZHANG, H., ZHANG, X., ZHAO, N., ZHANG, M., CHEN, M., ZHANG, S., XU, H., ZHAO, L., QIAO, M. and WU, Z. (2022). Bacteroides acidifaciens in the gut plays a protective role against CD95-mediated liver injury. *Gut Microbes* **14**.

WANG, L., CHEN, G. and LI, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23** 1486–1494.

WANG, L., WU, Y. and LI, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *J*. *Amer*. *Statist*. *Assoc*. **107** 214–222. MR2949353 https://doi.org/10.1080/01621459.2012.656014

WILLIAMS, J., BRAVO, H. C., TOM, J. and PAULSON, J. N. (2020). microbiomeDASim: simulating longitudinal differential abundance for microbiome data. *F*1000*Res*. **8** 1769.

XIA, F., CHEN, J., FUNG, W. K. and LI, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69** 1053–1063. MR3146800 https://doi.org/10.1111/biom.12079

YANG, Y. and ZOU, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Stat*. *Comput*. **25** 1129–1141. MR3401877 https://doi.org/10.1007/s11222-014-9498-5

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann*. *Statist*. **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHANG, X., MALLICK, H., TANG, Z., ZHANG, L., CUI, X., BENSON, A. K. and YI, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinform*. **18**.

ZHANG, X., PEI, Y.-F., ZHANG, L., GUO, B., PENDEGRAFT, A. H., ZHUANG, W. and YI, N. (2018). Negative binomial mixed models for analyzing longitudinal microbiome data. *Front*. *Microbiol*. **9**.

ZHANG, X. and YI, N. (2020). Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data. *Bioinformatics* **36** 2345–2351. https://doi.org/10.1093/bioinformatics/btz973

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

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann*. *Statist*. **36** 1509–1533. MR2435443 https://doi.org/10.1214/009053607000000802

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 967–985
https://doi.org/10.1214/24-AOAS1999

A UNIFIED QUANTILE FRAMEWORK FOR NONLINEAR HETEROGENEOUS TRANSCRIPTOME-WIDE ASSOCIATIONS

BY TIANYING WANG[1,a], IULIANA IONITA-LAZA[2,3,b] AND YING WEI[2,c]

[1]*Department of Statistics, Colorado State University,* [a]*Tianying.Wang@colostate.edu*
[2]*Department of Biostatistics, Columbia University,* [b]*ii2135@cumc.columbia.edu,* [c]*yw2148@cumc.columbia.edu*
[3]*Department of Statistics, Lund University*

Transcriptome-wide association studies (TWAS) are powerful tools for identifying gene-level associations by integrating genome-wide association studies and gene expression data. However, most TWAS methods focus on linear associations between genes and traits, ignoring the complex nonlinear relationships that may be present in biological systems. To address this limitation, we propose a novel framework, QTWAS, which integrates a quantile-based gene expression model into the TWAS model, allowing for the discovery of nonlinear and heterogeneous gene-trait associations. Via comprehensive simulations and applications to both continuous and binary traits, we demonstrate that the proposed model is more powerful than conventional TWAS in identifying gene-trait associations.

REFERENCES

BARBEIRA, A. N., DICKINSON, S. P., BONAZZOLA, R., ZHENG, J., WHEELER, H. E., TORRES, J. M., TORSTENSON, E. S., SHAH, K. P., GARCIA, T. et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. *Nat. Commun.* **9** 1825.

BARFIELD, R., FENG, H., GUSEV, A., WU, L., ZHENG, W., PASANIUC, B. and KRAFT, P. (2018). Transcriptome-wide association studies accounting for colocalization using egger regression. *Genet. Epidemiol.* **42** 418–433.

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

BUDINSKA, E., POPOVICI, V., TEJPAR, S., D'ARIO, G., LAPIQUE, N., SIKORA, K. O., DI NARZO, A. F., YAN, P., HODGSON, J. G. et al. (2013). Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J. Pathol.* **231** 63–76.

DENG, Y. and PAN, W. (2021). Model checking via testing for direct effects in Mendelian randomization and transcriptome-wide association studies. *PLoS Comput. Biol.* **17** e1009266.

DEVLIN, B. and ROEDER, K. (1999). Genomic control for association studies. *Biometrics* **55** 997–1004.

DONG, X., SU, Y.-R., BARFIELD, R., BIEN, S. A., HE, Q., HARRISON, T. A., HUYGHE, J. R., KEKU, T. O., LINDOR, N. M. et al. (2020). A general framework for functionally informed set-based analysis: Application to a large-scale colorectal cancer study. *PLoS Genet.* **16** e1008947.

DUDOIT, S., SHAFFER, J. P. and BOLDRICK, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18** 71–103. MR1997066 https://doi.org/10.1214/ss/1056397487

GAMAZON, E. R., WHEELER, H. E., SHAH, K. P., MOZAFFARI, S. V., AQUINO-MICHAELS, K., CARROLL, R. J., EYLER, A. E., DENNY, J. C., NICOLAE, D. L. et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47** 1091.

GIAMBARTOLOMEI, C., VUKCEVIC, D., SCHADT, E. E., FRANKE, L., HINGORANI, A. D., WALLACE, C. and PLAGNOL, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10** e1004383.

GLEASON, K. J., YANG, F. and CHEN, L. S. (2021). A robust two-sample transcriptome-wide Mendelian randomization method integrating gwas with multi-tissue eqtl summary statistics. *Genet. Epidemiol.* **45** 353–371.

GRAHAM, S. E., CLARKE, S. L., WU, K.-H. H., KANONI, S., ZAJAC, G. J., RAMDAS, S., SURAKKA, I., NTALLA, I., VEDANTAM, S. et al. (2021). The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600** 675–679.

Key words and phrases. Regression quantile process, nonlinear association test, transcriptome-wide association studies.

GTEx Consortium (2020a). The gtex consortium atlas of genetic regulatory effects across human tissues. *Science* **369** 1318–1330.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma, D. I. et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat*. *Genet*. **48** 245.

Gutenbrunner, C., Jurečková, J., Koenker, R. and Portnoy, S. (1993). Tests of linear hypotheses based on regression rank scores. *J*. *Nonparametr*. *Stat*. **2** 307–331. MR1256383 https://doi.org/10.1080/10485259308832561

Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., Yu, Z., Li, B., Gu, J. et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *bioRxiv*, 286013.

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

Koenker, R. and Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *J*. *Amer*. *Statist*. *Assoc*. **94** 1296–1310. MR1731491 https://doi.org/10.2307/2669943

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. **3** e161.

Li, B., Veturi, Y., Verma, A., Bradford, Y., Daar, E. S., Gulick, R. M., Riddler, S. A., Robbins, G. K., Lennox, J. L. et al. (2021). Tissue specificity-aware twas (tsa-twas) framework identifies novel associations with metabolic, immunologic, and virologic traits in hiv-positive adults. *PLoS Genet*. **17** e1009464.

Lin, Z., Xue, H., Malakhov, M. M., Knutson, K. A. and Pan, W. (2022). Accounting for nonlinear effects of gene expression identifies additional associated genes in transcriptomewide association studies. *Human molecular genetics*.

Liu, Y. and Xie, J. (2020). Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *J*. *Amer*. *Statist*. *Assoc*. **115** 393–402. MR4078471 https://doi.org/10.1080/01621459.2018.1554485

Ma, S., Dalgleish, J., Lee, J., Wang, C., Liu, L., Gill, R., Buxbaum, J. D., Chung, W. K., Aschard, H. et al. (2021). Powerful gene-based testing by integrating long-range chromatin interactions and knockoff genotypes. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **118** e2105191118.

Ma, S., Wang, C., Khan, A., Liu, L., Dalgleish, J., Kiryluk, K., He, Z. and Ionita-Laza, I. (2023). Bigknock: Fine-mapping gene-based associations via knockoff analysis of biobank-scale data. *Genome Biol*. **24** 24.

Mancuso, N., Freund, M. K., Johnson, R., Shi, H., Kichaev, G., Gusev, A. and Pasaniuc, B. (2019). Probabilistic fine-mapping of transcriptome-wide association studies. *Nat*. *Genet*. **51** 675–682.

Nagpal, S., Meng, X., Epstein, M. P., Tsoi, L. C., Patrick, M., Gibson, G., De Jager, P. L., Bennett, D. A., Wingo, A. P. et al. (2019). Tigar: An improved Bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. *Am*. *J*. *Hum*. *Genet*. **105** 258–266.

O'Connor, L. J., Gusev, A., Liu, X., Loh, P.-R., Finucane, H. K. and Price, A. L. (2017). Estimating the proportion of disease heritability mediated by gene expression levels. *BioRxiv* **118018**.

Pan-UKB Team (2020b). https://pan.ukbb.broadinstitute.org.

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S. E., Bishop, S., Cameron, D. et al. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat*. *Genet*. **50** 381–389.

Plagnol, V., Smyth, D. J., Todd, J. A. and Clayton, D. G. (2009). Statistical independence of the colocalized association signals for type 1 diabetes and rps26 gene expression on chromosome 12q13. *Biostatistics* **10** 327–334.

PredictDB Team (2021). Gtex v8 models on eqtl and sqtl. /post/2021/07/21/gtex-v8-models-on-eqtl-and-sqtl/.

Shi, X., Chai, X., Yang, Y., Cheng, Q., Jiao, Y., Chen, H., Huang, J., Yang, C. and Liu, J. (2020). A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies. *Nucleic Acids Res*. **48** e109–e109.

Somel, M., Khaitovich, P., Bahn, S., Pääbo, S. and Lachmann, M. (2006). Gene expression becomes heterogeneous with age. *Curr*. *Biol*. **16** R359–R360.

Song, X., Li, G., Zhou, Z., Wang, X., Ionita-Laza, I. and Wei, Y. (2017). Qrank: A novel quantile regression tool for eqtl discovery. *Bioinformatics* **33** 2123–2130.

Stegle, O., Parts, L., Durbin, R. and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput*. *Biol*. **6** e1000770. MR2659818 https://doi.org/10.1371/journal.pcbi.1000770

Stegle, O., Parts, L., Piipari, M., Winn, J. and Durbin, R. (2012). Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nat*. *Protoc*. **7** 500–507.

TANG, S., BUCHMAN, A. S., DE JAGER, P. L., BENNETT, D. A., EPSTEIN, M. P. and YANG, J. (2021). Novel variance-component twas method for studying complex human diseases with applications to Alzheimer's dementia. *PLoS Genet*. **17** e1009482.

UMANS, B. D., BATTLE, A. and GILAD, Y. (2021). Where are the disease-associated eqtls? *Trends Genet*. **37** 109–124.

VAN DER GRAAF, A., WARMERDAM, R., AUWERX, C. M. P., VOSA, U., BORGES, M. C., FRANKE, L. and KUTALIK, Z. (2024). Mr-link-2: Pleiotropy robust cis Mendelian randomization validated in four independent gold-standard datasets of causality. *medRxiv*, 2024–01.

WAINBERG, M., SINNOTT-ARMSTRONG, N., MANCUSO, N., BARBEIRA, A. N., KNOWLES, D. A., GOLAN, D., ERMEL, R., RUUSALEPP, A., QUERTERMOUS, T. et al. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet*. **51** 592–599.

WANG, C., WANG, T., KIRYLUK, K., WEI, Y., ASCHARD, H. and IONITA-LAZA, I. (2024). Genome-wide discovery for biomarkers using quantile regression at biobank scale. *Nat. Commun*. **15** 6460.

WANG, T., IONITA-LAZA, I. and WEI, Y. (2022). Integrated Quantile RAnk Test (iQRAT) for gene-level associations. *Ann. Appl. Stat*. **16** 1423–1444. MR4455887 https://doi.org/10.1214/21-aoas1548

WANG, T., IONITA-LAZA, I. and WEI, Y. (2025). Supplement to "A unified quantile framework for nonlinear heterogeneous transcriptome-wide associations." https://doi.org/10.1214/24-AOAS1999SUPPA, https://doi.org/10.1214/24-AOAS1999SUPPB

ZHAO, B., SHAN, Y., YANG, Y., YU, Z., LI, T., WANG, X., LUO, T., ZHU, Z., SULLIVAN, P. et al. (2021). Transcriptome-wide association analysis of brain structures yields insights into pleiotropy with complex neuropsychiatric traits. *Nat. Commun*. **12** 1–11.

ZHAO, S., CROUSE, W., QIAN, S., LUO, K., STEPHENS, M. and HE, X. (2024). Adjusting for genetic confounders in transcriptome-wide association studies improves discovery of risk genes of complex traits. *Nat. Genet*. 1–12.

ZHU, H. and ZHOU, X. (2021). Transcriptome-wide association studies: A view from Mendelian randomization. *Quant. Biol*. **9** 107–121.

ZUBER, V., GRINBERG, N. F., GILL, D., MANIPUR, I., SLOB, E. A., PATEL, A., WALLACE, C. and BURGESS, S. (2022). Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *Am. J. Hum. Genet*. **109** 767–782.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 986–1005
https://doi.org/10.1214/24-AOAS2001

CLUSTER ANALYSIS OF LONGITUDINAL PROFILES FOR COMPOSITIONAL COUNT DATA

BY CHENYANG DUAN[1,a] AND YUAN JIANG[2,b]

[1]*Data and Statistical Sciences, AbbVie,* [a]*chenyang.duan@abbvie.com*
[2]*Department of Statistics, Oregon State University,* [b]*yuan.jiang@oregonstate.edu*

To classify biological roles of different species in an ecological system, modern studies collect longitudinal and compositional counts of DNA sequences of taxonomically diagnostic genetic markers to measure the abundance of species over time. The major challenges of conducting this analysis are twofold: how to accommodate the complex dependence in this data type and how to model the longitudinal trajectories of the species' abundances. In this paper we propose a novel method named COMPARING to cluster longitudinal profiles for compositional count data to address these challenges. In COMPARING, generalized estimating equation is used to account for both the compositional and longitudinal dependence structures, nonparametric B-spline approximation is used to model the longitudinal curves, and a pairwise-distance penalization is used to identify subgroups with similar longitudinal patterns. We establish the convergence rate of the estimated curves and conclude that the true subgroups can be correctly identified with a high probability. We also conduct simulation studies to show the advantage of COMPARING over its competitors in clustering longitudinal trajectories from compositional count data. Finally, we apply COMPARING to study the coexistence of blood-borne parasites in African buffalo and demonstrate how the method successfully detects biologically meaningful subgroups of parasites for competition-colonization trade-off.

REFERENCES

ABRAHAM, C., CORNILLON, P. A., MATZNER-LØBER, E. and MOLINARI, N. (2003). Unsupervised curve clustering using B-splines. *Scand. J. Stat.* **30** 581–595. MR2002229 https://doi.org/10.1111/1467-9469.00350

CHEN, B. and XU, W. (2020). Generalized estimating equation modeling on correlated microbiome sequencing data with longitudinal measures. *PLoS Comput. Biol.* **16** e1008108.

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

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

COFFEY, N., HINDE, J. and HOLIAN, E. (2014). Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data. *Comput. Statist. Data Anal.* **71** 14–29. MR3131951 https://doi.org/10.1016/j.csda.2013.04.001

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

DUAN, C. and JIANG, Y. (2025). Supplement to "Cluster Analysis of Longitudinal Profiles for Compositional Count Data." https://doi.org/10.1214/24-AOAS2001SUPP

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

ESSER, E., ZHANG, X. and CHAN, T. F. (2010). A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM J. Imaging Sci.* **3** 1015–1046. MR2763706 https://doi.org/10.1137/09076934X

GABAY, D. and MERCIER, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.* **2** 17–40. https://doi.org/10.1016/0898-1221(76)90003-1

Key words and phrases. B-spline, competition-colonization trade-off, generalized estimating equation, linearized alternating direction method of multipliers, working correlation structure.

GLIDDEN, C. K., KARAKOÇ, C., DUAN, C., JIANG, Y., BEECHLER, B., JABBAR, A. and JOLLES, A. E. (2023). Distinct life history strategies underpin clear patterns of succession in microparasite communities infecting a wild mammalian host. *Mol. Ecol.* **32** 3733–3746.

HASTINGS, A. (1980). Disturbance, coexistence, history, and competition for space. *Theor. Popul. Biol.* **18** 363–373. MR0620634 https://doi.org/10.1016/0040-5809(80)90059-3

HESTENES, M. R. (1969). Multiplier and gradient methods. *J. Optim. Theory Appl.* **4** 303–320. MR0271809 https://doi.org/10.1007/BF00927673

LEVIN, S. A. (1970). Community equilibria and stability, and an extension of the competitive exclusion principle. *Amer. Nat.* **104** 413–423.

LEVINS, R. and CULVER, D. (1971). Regional coexistence of species and competition between rare species. *Proc. Natl. Acad. Sci. USA* **68** 1246–1248.

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430 https://doi.org/10.1093/biomet/73.1.13

LIU, Q., SHEN, X. and GU, Y. (2019). Linearized ADMM for nonconvex nonsmooth optimization with convergence analysis. *IEEE Access* **7** 76131–76144.

LUAN, Y. and LI, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics* **19** 474–482.

MA, P., CASTILLO-DAVIS, C. I., ZHONG, W. and LIU, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Res.* **34** 1261–1269.

MA, S. and HUANG, J. (2017). A concave pairwise fusion approach to subgroup analysis. *J. Amer. Statist. Assoc.* **112** 410–423. MR3646581 https://doi.org/10.1080/01621459.2016.1148039

PAN, W., SHEN, X. and LIU, B. (2013). Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *J. Mach. Learn. Res.* **14** 1865–1889. MR3104498

XIA, F., CHEN, J., FUNG, W. K. and LI, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69** 1053–1063. MR3146800 https://doi.org/10.1111/biom.12079

YANG, Y., CHEN, N. and CHEN, T. (2017). Inference of environmental factor-microbe and microbe-microbe associations from metagenomic data using a hierarchical Bayesian statistical model. *Cell Syst.* **4** 129–137.

YOSHIDA, T. and NAITO, K. (2012). Asymptotics for penalized additive B-spline regression. *J. Japan Statist. Soc.* **42** 81–107. MR3051878 https://doi.org/10.14490/jjss.42.81

YOSHIDA, T. and NAITO, K. (2014). Asymptotics for penalised splines in generalised additive models. *J. Nonparametr. Stat.* **26** 269–289. MR3195983 https://doi.org/10.1080/10485252.2014.899360

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHU, X. and QU, A. (2018). Cluster analysis of longitudinal profiles with subgroups. *Electron. J. Stat.* **12** 171–193. MR3756096 https://doi.org/10.1214/17-EJS1389

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1006–1027
https://doi.org/10.1214/25-AOAS2012

AUGMENTED DOUBLY ROBUST POST-IMPUTATION INFERENCE FOR PROTEOMIC DATA

BY HAEUN MOON[1,a], JIN-HONG DU[2,b], JING LEI[2,c] AND KATHRYN ROEDER[2,d]

[1]*Department of Statistics, Seoul National University,* [a]*haeunmoon@snu.ac.kr*
[2]*Department of Statistics and Data Science, Carnegie Mellon University,* [b]*jinhongd@andrew.cmu.edu,* [c]*jinglei@andrew.cmu.edu,* [d]*roeder@andrew.cmu.edu*

Quantitative measurements produced by mass spectrometry proteomics experiments offer a direct way to explore the role of proteins in molecular mechanisms. However, analysis of such data is challenging due to the large proportion of missing values. A common strategy to address this issue is to utilize an imputed dataset, which often introduces systematic bias into downstream analyses if the imputation errors are ignored. In this paper we propose a statistical framework, inspired by doubly robust estimators, that offers valid and efficient inference for proteomic data. Our framework combines powerful machine learning tools, such as variational autoencoders, to augment the imputation quality with high-dimensional peptide data, and a parametric model to estimate the propensity score for debiasing imputed outcomes. Our estimator is compatible with the double machine learning framework and has provable properties. Simulation studies verify its empirical superiority over other existing procedures. In application to both single-cell proteomic data and bulk-cell Alzheimer's disease data our method utilizes the imputed data to gain additional, meaningful discoveries and yet maintains good control of false positives.

REFERENCES

AGARWAL, D., WANG, J. and ZHANG, N. R. (2020). Data denoising and post-denoising corrections in single cell RNA sequencing. *Statist. Sci.* **35** 112–128. MR4071361 https://doi.org/10.1214/19-STS7560

ANDREWS, T. S. and HEMBERG, M. (2018). False signals induced by single-cell imputation. *F1000Res.* **7** 1740. https://doi.org/10.12688/f1000research.16613.2

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

BERREVOETS, J., IMRIE, F., KYONO, T., JORDON, J. and VAN DER SCHAAR, M. (2023). To impute or not to impute? Missing data in treatment effect estimation. In *International Conference on Artificial Intelligence and Statistics* 3568–3590. PMLR.

BHATIA, S., RAWAL, R., SHARMA, P., SINGH, T., SINGH, M. and SINGH, V. (2022). Mitochondrial dysfunction in Alzheimer's disease: Opportunities for drug development. *Curr. Neuropharmacology* **20** 675.

BRAMER, L. M., IRVAHN, J., PIEHOWSKI, P. D., RODLAND, K. D. and WEBB-ROBERTSON, B.-J. M. (2021). A review of imputation strategies for isobaric labeling-based shotgun proteomics. *J. Proteome Res.* **20** 1–13. https://doi.org/10.1021/acs.jproteome.0c00123

BRENES, A., HUKELMANN, J., BENSADDEK, D. and LAMOND, A. I. (2019). Multibatch TMT reveals false positives, batch effects and missing values. *Mol. Cell. Proteomics* **18** 1967–1980. https://doi.org/10.1074/mcp.RA119.001472

BRINI, A. and VAN DEN HEUVEL, E. R. (2024). Missing data imputation with high-dimensional data. *Amer. Statist.* **78** 240–252. MR4734074 https://doi.org/10.1080/00031305.2023.2259962

CHEN, L. S., WANG, J., WANG, X. and WANG, P. (2017). A mixed-effects model for incomplete data from labeling-based quantitative proteomics experiments. *Ann. Appl. Stat.* **11** 114–138. MR3634317 https://doi.org/10.1214/16-AOAS994

CHEN, Y., FAN, J., MA, C. and YAN, Y. (2019). Inference and uncertainty quantification for noisy matrix completion. *Proc. Natl. Acad. Sci. USA* **116** 22931–22937. MR4036123 https://doi.org/10.1073/pnas.1910053116

Key words and phrases. Proteomic data, post-imputation inference, double robustness, variational autoencoder.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CHION, M., CARAPITO, C. and BERTRAND, F. (2022). Accounting for multiple imputation-induced variability for differential analysis in mass spectrometry-based label-free quantitative proteomics. *PLoS Comput. Biol.* **18** e1010420.

DAI, C., LIN, B., XING, X. and LIU, J. S. (2023). A scale-free approach for false discovery rate control in generalized linear models. *J. Amer. Statist. Assoc.* **118** 1551–1565. MR4646583 https://doi.org/10.1080/01621459.2023.2165930

DAVIDIAN, M. (2022). Methods based on semiparametric theory for analysis in the presence of missing data. *Annu. Rev. Stat. Appl.* **9** 167–196. MR4394905 https://doi.org/10.1146/annurev-statistics-040120-025906

DELHAYE, S., JARJAT, M., BOULKSIBAT, A., SANCHEZ, C., TEMPIO, A., TURTOI, A., GIORGI, M., LACAS-GERVAIS, S., BAJ, G. et al. (2024). Defects in AMPAR trafficking and microglia activation underlie socio-cognitive deficits associated to decreased expression of phosphodiesterase 2 a. *Neurobiol. Dis.* **191** 106393. https://doi.org/10.1016/j.nbd.2023.106393

DÍAZ, I., SAVENKOV, O. and BALLMAN, K. (2018). Targeted learning ensembles for optimal individualized treatment rules with time-to-event outcomes. *Biometrika* **105** 723–738. MR3842895 https://doi.org/10.1093/biomet/asy017

DILEEP, V., BOIX, C. A., MATHYS, H., MARCO, A., WELCH, G. M., MEHARENA, H. S., LOON, A., JELOKA, R., PENG, Z. et al. (2023). Neuronal DNA double-strand breaks lead to genome structural variations and 3D genome disruption in neurodegeneration. *Cell* **186** 4404–4421.

DU, J.-H., CAI, Z. and ROEDER, K. (2022). Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scVAEIT. *Proc. Natl. Acad. Sci. USA* **119** e2214414119. https://doi.org/10.1073/pnas.2214414119

DU, L., GUO, X., SUN, W. and ZOU, C. (2023). False discovery rate control under general dependence by symmetrized data aggregation. *J. Amer. Statist. Assoc.* **118** 607–621. MR4571145 https://doi.org/10.1080/01621459.2021.1945459

FISHER, A. and FISHER, V. (2023). Three-way cross-fitting and pseudo-outcome regression for estimation of conditional effects and other linear functionals. arXiv preprint. Available at arXiv:2306.07230.

GIANETTO, Q. G., WIECZOREK, S., COUTÉ, Y. and BURGER, T. (2020). A peptide-level multiple imputation strategy accounting for the different natures of missing values in proteomics data. *BioRxiv* 2020–05.

GOURIÉROUX, C. and MONFORT, A. (1981). Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *J. Econometrics* **17** 83–97. MR0640525 https://doi.org/10.1016/0304-4076(81)90060-9

GUI, Y., BARBER, R. and MA, C. (2023). Conformalized matrix completion. *Adv. Neural Inf. Process. Syst.* **36** 4820–4844.

GUO, X., REN, H., ZOU, C. and LI, R. (2023). Threshold selection in feature screening for error rate control. *J. Amer. Statist. Assoc.* **118** 1773–1785. MR4646605 https://doi.org/10.1080/01621459.2021.2011735

HARRIS, L., FONDRIE, W. E., OH, S. and NOBLE, W. S. (2023). Evaluating proteomics imputation methods with improved criteria. *J. Proteome Res.* **22** 3427–3438. https://doi.org/10.1021/acs.jproteome.3c00205

HASTIE, T., MAZUMDER, R., LEE, J. D. and ZADEH, R. (2015). Matrix completion and low-rank SVD via fast alternating least squares. *J. Mach. Learn. Res.* **16** 3367–3402. MR3450542

HIGHAM, J. P., MALIK, B. R., BUHL, E., DAWSON, J. M., OGIER, A. S., LUNNON, K. and HODGE, J. J. L. (2019). Alzheimer's disease associated genes ankyrin and tau cause shortened lifespan and memory loss in drosophila. *Front. Cell. Neurosci.* **13** 260.

INUKAI, R., MORI, K., KUWATA, K., SUZUKI, C., MAKI, M., TAKAHARA, T. and SHIBATA, H. (2021). The novel ALG-2 target protein CDIP1 promotes cell death by interacting with ESCRT-I and VAPA/B. *Int. J. Mol. Sci.* **22**. https://doi.org/10.3390/ijms22031175

JIANG, K., MUKHERJEE, R., SEN, S. and SUR, P. (2022). A new central limit theorem for the augmented IPW estimator: Variance inflation, cross-fit covariance and beyond. arXiv preprint. Available at arXiv:2205.10198.

JONES, I., DENT, L., HIGO, T., ROUMELIOTIS, T., ARIAS GARCIA, M., SHREE, H., CHOUDHARY, J., PEDERSEN, M. and BAKAL, C. (2023). Characterization of proteome-size scaling by integrative omics reveals mechanisms of proliferation control in cancer. *Sci. Adv.* **9** eadd0636.

KENNEDY, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electron. J. Stat.* **17** 3008–3049. MR4667730 https://doi.org/10.1214/23-ejs2157

KUMARI, A., RAHAMAN, A., ZENG, X.-A., FAROOQ, M. A., HUANG, Y., YAO, R., ALI, M., ISHRAT, R. and ALI, R. (2022). Temporal cortex microarray analysis revealed impaired ribosomal biogenesis and hyperactivity of the glutamatergic system: An early signature of asymptomatic Alzheimer's disease. *Front. Neurosci.* **16** 966877.

LANZ, M. C., ZATULOVSKIY, E., SWAFFER, M. P., ZHANG, L., ILERTEN, I., ZHANG, S., YOU, D. S., MARINOV, G., MCALPINE, P. et al. (2022). Increasing cell size remodels the proteome and promotes senescence. *Mol. Cell* **82** 3255–3269.

LEDUC, A., HUFFMAN, R. G., CANTLON, J., KHAN, S. and SLAVOV, N. (2022). Exploring functional protein covariation across single cells using nPOP. *Genome Biol.* **23** 261.

LEE, J. D., SUN, Y. and SAUNDERS, M. A. (2014). Proximal Newton-type methods for minimizing composite functions. *SIAM J. Optim.* **24** 1420–1443. MR3252813 https://doi.org/10.1137/130921428

LENGEFELD, J., CHENG, C.-W., MARETICH, P., BLAIR, M., HAGEN, H., MCREYNOLDS, M. R., SULLIVAN, E., MAJORS, K., ROBERTS, C. et al. (2021). Cell size is a determinant of stem cell potential during aging. *Sci. Adv.* **7** eabk0271.

LITTLE, R. J. (1992). Regression with missing X's: A review. *J. Amer. Statist. Assoc.* **87** 1227–1237.

LITTLE, R. J., D'AGOSTINO, R., COHEN, M. L., DICKERSIN, K., EMERSON, S. S., FARRAR, J. T., FRANGAKIS, C., HOGAN, J. W., MOLENBERGHS, G. et al. (2012). The prevention and treatment of missing data in clinical trials. *N. Engl. J. Med.* **367** 1355–1360.

LIU, M. and DONGRE, A. (2021). Proper imputation of missing values in proteomics datasets for differential expression analysis. *Brief. Bioinform.* **22**. https://doi.org/10.1093/bib/bbaa112

LIU, X., YAN, J. and KIRSCHNER, M. W. (2024). Cell size homeostasis is tightly controlled throughout the cell cycle. *PLoS Biol.* **22** e3002453. https://doi.org/10.1371/journal.pbio.3002453

LIU, Y., BEYER, A. and AEBERSOLD, R. (2016). On the dependency of cellular protein levels on mRNA abundance. *Cell* **165** 535–550. https://doi.org/10.1016/j.cell.2016.03.014

LLOYD, A. C. (2013). The regulation of cell size. *Cell* **154** 1194–1205. https://doi.org/10.1016/j.cell.2013.08.053

LY, L.-H. and VINGRON, M. (2022). Effect of imputation on gene network reconstruction from single-cell RNA-seq data. *Patterns* **3** 100414. https://doi.org/10.1016/j.patter.2021.100414

MA, S. L., HUANG, W., TANG, N. L. and LAM, L. C. (2012). MxA polymorphisms are associated with risk and age-at-onset in Alzheimer disease and accelerated cognitive decline in Chinese elders. *Rejuvenation Res.* **15** 516–522.

MACDONALD, M. L., ALHASSAN, J., NEWMAN, J. T., RICHARD, M., GU, H., KELLY, R. M., SAMPSON, A. R., FISH, K. N., PENZES, P. et al. (2017). Selective loss of smaller spines in schizophrenia. *Am. J. Psychiatr.* **174** 586–594. https://doi.org/10.1176/appi.ajp.2017.16070814

MARGUERAT, S. and BÄHLER, J. (2012). Coordinating genome expression with cell size. *Trends Genet.* **28** 560–565. https://doi.org/10.1016/j.tig.2012.07.003

MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* 538–558.

MERRIHEW, G. E., PARK, J., PLUBELL, D., SEARLE, B. C., KEENE, C. D., LARSON, E. B., BATEMAN, R., PERRIN, R. J., CHHATWAL, J. P. et al. (2023). A peptide-centric quantitative proteomics dataset for the phenotypic assessment of Alzheimer's disease. *Sci. Data* **10** 206.

MOLL, T., SHAW, P. J. and COOPER-KNOCK, J. (2020). Disrupted glycosylation of lipids and proteins is a cause of neurodegeneration. *Brain* **143** 1332–1340. https://doi.org/10.1093/brain/awz358

MOON, H., DU, J.-H., LEI, J. and ROEDER, K. (2025). Supplement to "Augmented doubly robust post-imputation inference for proteomic data." https://doi.org/10.1214/25-AOAS2012SUPPA, https://doi.org/10.1214/25-AOAS2012SUPPB

PEREIRA, A. C., GRAY, J. D., KOGAN, J. F., DAVIDSON, R. L., RUBIN, T. G., OKAMOTO, M., MORRISON, J. H. and MCEWEN, B. S. (2017). Age and Alzheimer's disease gene expression profiles reversed by the glutamate modulator riluzole. *Mol. Psychiatry* **22** 296–305. https://doi.org/10.1038/mp.2016.33

PRAKASH, P., ERDJUMENT-BROMAGE, H., O'DEA, M. R., MUNSON, C. N., LABIB, D., FOSSATI, V., NEUBERT, T. A. and LIDDELOW, S. A. (2024). Proteomic profiling of interferon-responsive reactive astrocytes in rodent and human. *GLIA* **72** 625–642. https://doi.org/10.1002/glia.24494

QIU, Y. and MESSER, K. (2023). An efficient doubly-robust imputation framework for longitudinal dropout, with an application to an Alzheimer's clinical trial. *Ann. Appl. Stat.* **17** 2473–2493. MR4637676 https://doi.org/10.1214/23-aoas1728

QIU, Y. L., ZHENG, H. and GEVAERT, O. (2020). Genomic data imputation with variational auto-encoders. *GigaScience* **9** giaa082.

RASHID, M. and SHIFA, N. (2009). Consistency of the maximum likelihood estimator in logistic regression model: A different approach. *J. Stat.* **16**.

ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90** 122–129. MR1325119

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

RYCKMAN, A. E., BROCKHAUSEN, I. and WALIA, J. S. (2020). Metabolism of glycosphingolipids and their role in the pathophysiology of lysosomal storage disorders. *Int. J. Mol. Sci.* **21**. https://doi.org/10.3390/ijms21186881

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478 https://doi.org/10.2307/2669923

SEMENOVA, V. and CHERNOZHUKOV, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *Econom. J.* **24** 264–289. MR4281225 https://doi.org/10.1093/ectj/utaa027

SHAO, M. and ZHANG, Y. (2023). Distribution-free matrix prediction under arbitrary missing pattern. arXiv preprint. Available at arXiv:2305.11640.

SHEN, M., CHANG, Y.-T., WU, C.-T., PARKER, S. J., SAYLOR, G., WANG, Y., YU, G., VAN EYK, J. E., CLARKE, R. et al. (2022). Comparative assessment and novel strategy on methods for imputing proteomics data. *Sci. Rep.* **12** 1067.

SHENG, J., ZHANG, S., WU, L., KUMAR, G., LIAO, Y., GK, P. and FAN, H. (2022). Inhibition of phosphodiesterase: A novel therapeutic target for the treatment of mild cognitive impairment and Alzheimer's disease. *Front. Aging Neurosci.* **14** 1019187.

SHI, J., LI, Y., ZHANG, Y., CHEN, J., GAO, J., ZHANG, T., SHANG, X. and ZHANG, X. (2021). Baicalein ameliorates Aβ-induced memory deficits and neuronal atrophy via Inhibition of PDE2 and PDE4. *Front. Pharmacol.* **12** 794458.

STEKHOVEN, D. J. and BÜHLMANN, P. (2012). MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28** 112–118. https://doi.org/10.1093/bioinformatics/btr597

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

TANG, X., TENA, J., LUCENTE, J. D., MAEZAWA, I., HARVEY, D. J., JIN, L.-W., LEBRILLA, C. B. and ZIVKOVIC, A. M. (2023). Transcriptomic and glycomic analyses highlight pathway-specific glycosylation alterations unique to Alzheimer's disease. *Sci. Rep.* **13** 7816. https://doi.org/10.1038/s41598-023-34787-4

TASAKI, S., XU, J., AVEY, D. R., JOHNSON, L., PETYUK, V. A., DAWE, R. J., BENNETT, D. A., WANG, Y. and GAITERI, C. (2022). Inferring protein expression changes from mRNA in Alzheimer's dementia using deep neural networks. *Nat. Commun.* **13** 655.

TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., BROWN, P., HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D. and ALTMAN, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* **17** 520–525.

TURNER, J. J., EWALD, J. C. and SKOTHEIM, J. M. (2012). Cell size control in yeast. *Curr. Biol.* **22** R350–R359. https://doi.org/10.1016/j.cub.2012.02.041

VÄLIKANGAS, T., SUOMI, T. and ELO, L. L. (2018). A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. *Brief. Bioinform.* **19** 1344–1355.

VANDERAA, C. and GATTO, L. (2023). Revisiting the thorny issue of missing values in single-cell proteomics. *J. Proteome Res.* **22** 2775–2784. https://doi.org/10.1021/acs.jproteome.3c00227

VAN DER VAART, A. W. (2000). *Asymptotic Statistics* **3**. Cambridge University Press, Cambridge.

VOGEL, C. and MARCOTTE, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13** 227–232. https://doi.org/10.1038/nrg3185

WANG, J., AGARWAL, D., HUANG, M., HU, G., ZHOU, Z., YE, C. and ZHANG, N. R. (2019). Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods* **16** 875–878.

WANG, J., GAMAZON, E. R., PIERCE, B. L., STRANGER, B. E., IM, H. K., GIBBONS, R. D., COX, N. J., NICOLAE, D. L. and CHEN, L. S. (2016). Imputing gene expression in uncollected tissues within and beyond GTEx. *Am. J. Hum. Genet.* **98** 697–708.

WEBB-ROBERTSON, B.-J. M., WIBERG, H. K., MATZKE, M. M., BROWN, J. N., WANG, J., MCDERMOTT, J. E., SMITH, R. D., RODLAND, K. D., METZ, T. O. et al. (2015). Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J. Proteome Res.* **14** 1993–2001.

WEI, R., WANG, J., SU, M., JIA, E., CHEN, S., CHEN, T. and NI, Y. (2018). Missing value imputation approach for mass spectrometry-based metabolomics data. *Sci. Rep.* **8** 663.

WESSELING, H., XU, B., WANT, E. J., HOLMES, E., GUEST, P., KARAYIORGOU, M., GOGOS, J. and BAHN, S. (2017). System-based proteomic and metabonomic analysis of the Df (16) A+/- mouse identifies potential miR-185 targets and molecular pathway alterations. *Mol. Psychiatry* **22** 384–395.

WIDJAYA, M. A., CHENG, Y.-J., KUO, Y.-M., LIU, C.-H., CHENG, W.-C. and LEE, S.-D. (2023). Transcriptomic analyses of exercise training in Alzheimer's disease cerebral cortex. *J. Alzheimer's Dis.* **93** 349–363.

YADLOWSKY, S. (2022). Explaining practical differences between treatment effect estimators with high dimensional asymptotics. arXiv preprint. Available at arXiv:2203.12538.

YIN, X., LEVY, D., WILLINGER, C., ADOURIAN, A. and LARSON, M. G. (2016). Multiple imputation and analysis for high-dimensional incomplete proteomics data. *Stat. Med.* **35** 1315–1326. MR3512814 https://doi.org/10.1002/sim.6800

YOON, J., JORDON, J. and SCHAAR, M. (2018). Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning* 5689–5698. PMLR.

ZHAO, A. and DING, P. (2024). To adjust or not to adjust? Estimating the average treatment effect in randomized experiments with missing covariates. *J. Amer. Statist. Assoc.* **119** 450–460. MR4713905 https://doi.org/10.1080/01621459.2022.2123814

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1028–1047
https://doi.org/10.1214/25-AOAS2014

BAYESIAN NONPARAMETRIC CLUSTERING WITH FEATURE SELECTION FOR SPATIALLY RESOLVED TRANSCRIPTOMICS DATA

BY BENCONG ZHU[1,a], GUANYU HU[2,c], LIN XU[3,d], XIAODAN FAN[1,b] AND QIWEI LI[4,e]

[1]*Department of Statistics, Chinese University of Hong Kong,* [a]*bencongzhu@link.cuhk.edu.hk,* [b]*xfan@cuhk.edu.hk*
[2]*Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston,* [c]*guanyu.hu@uth.tmc.edu*
[3]*Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center,* [d]*lin.xu@utsouthwestern.edu*
[4]*Department of Mathematical Sciences, University of Texas at Dallas,* [e]*qiwei.li@utdallas.edu*

The advent of next-generation sequencing-based spatially resolved transcriptomics (SRT) techniques has reshaped genomic studies by enabling high-throughput gene expression profiling while preserving spatial context. Nevertheless, inherent challenges associated with these new high-dimensional spatial data (e.g., zero inflation) pose obstacles to effective clustering, a fundamental problem in SRT data analysis. Current computational approaches often rely on heuristic data preprocessing and arbitrary cluster number specification, leading to information loss and, consequently, suboptimal downstream analysis. In response, we present BNPSpace, a novel Bayesian nonparametric spatial clustering framework that directly models SRT count data and automatically estimates the optimal number of spatial domains. BNPSpace facilitates the partitioning of the heterogeneous spatial domain while identifying a parsimonious set of discriminating genes among spatial domains, enhancing the interpretability of the findings. Additionally, it incorporates spatial information through a Markov random field prior model, encouraging a biologically meaningful partition pattern. We assess the performance of BNPSpace utilizing both simulated and real SRT data, demonstrating each innovations above are essential for the accurate identification of spatial domains. We also verified its scalability with large-scale SRT data. In the application to the human dorsolateral prefrontal cortex (DLPFC) 10x Visium data, comprising 4788 spots, BNPSpace outperforms existing methods by identifying more coherent spatial domain patterns. Furthermore, the discriminating genes identified by BNPSpace showed significant enrichment with odd ratio $= 1.877$ (p-value $= 0.00162$) in DLPFC gene sets validated by real biological experiments, underscoring its effectiveness in revealing biologically relevant insights.

REFERENCES

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

ALLEN, C., CHANG, Y., NEELON, B., CHANG, W., KIM, H. J., LI, Z., MA, Q. and CHUNG, D. (2023). A Bayesian multivariate mixture model for high throughput spatial transcriptomics. *Biometrics* **79** 1775–1787. MR4643955 https://doi.org/10.1111/biom.13727

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969

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

BECHT, E., MCINNES, L., HEALY, J., DUTERTRE, C.-A., KWOK, I. W., NG, L. G., GINHOUX, F. and NEWELL, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37** 38–44.

Key words and phrases. Feature selection, high-dimensional count data, Markov random field, mixture of finite mixtures, spatial clustering, zero-inflated Poisson mixture model.

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

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.

BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355. MR0362614

BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R. and LEFEBVRE, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008** P10008.

BROWN, P. J., VANNUCCI, M. and FEARN, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 627–641. MR1626005 https://doi.org/10.1111/1467-9868.00144

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

CHEN, K. H., BOETTIGER, A. N., MOFFITT, J. R., WANG, S. and ZHUANG, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348** aaa6090.

DAHL, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics* **4** 201–218.

DE BRUIN, E. C., MCGRANAHAN, N., MITTER, R., SALM, M., WEDGE, D. C., YATES, L., JAMAL-HANJANI, M., SHAFI, S., MURUGAESU, N. et al. (2014). Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346** 251–256.

FINAK, G., MCDAVID, A., YAJIMA, M., DENG, J., GERSUK, V., SHALEK, A. K., SLICHTER, C. K., MILLER, H. W., MCELRATH, M. J. et al. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16** 1–13.

GENG, J., BHATTACHARYA, A. and PATI, D. (2019). Probabilistic community detection with unknown number of communities. *J. Amer. Statist. Assoc.* **114** 893–905. MR3963189 https://doi.org/10.1080/01621459.2018.1458618

GENG, J., SHI, W. and HU, G. (2021). Bayesian nonparametric nonhomogeneous Poisson process with applications to USGS earthquake data. *Spat. Stat.* **41** Paper No. 100495, 26 pp. MR4214392 https://doi.org/10.1016/j.spasta.2021.100495

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.

GRUENE, M., PFLAUM, M., HESS, C., DIAMANTOUROS, S., SCHLIE, S., DEIWICK, A., KOCH, L., WILHELMI, M., JOCKENHOEVEL, S. et al. (2011). Laser printing of three-dimensional multicellular arrays for studies of cell-cell and cell–environment interactions. *Tissue Eng. Part C Methods* **17** 973–982.

HU, G., GENG, J., XUE, Y. and SANG, H. (2023). Bayesian spatial homogeneity pursuit of functional data: An application to the U.S. income distribution. *Bayesian Anal.* **18** 579–605. MR4578065 https://doi.org/10.1214/22-ba1320

HU, J., LI, X., COLEMAN, K., SCHROEDER, A., MA, N., IRWIN, D. J., LEE, E. B., SHINOHARA, R. T. and LI, M. (2021). SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18** 1342–1351.

JAIN, S. and NEAL, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *J. Comput. Graph. Statist.* **13** 158–182. MR2044876 https://doi.org/10.1198/1061860043001

JIANG, X., WANG, S., GUO, L., ZHU, B., WEN, Z., JIA, L., XU, L., XIAO, G. and LI, Q. (2024). iIMPACT: Integrating image and molecular profiles for spatial transcriptomics analysis. *Genome Biol.* **25** 1–25.

KISELEV, V. Y., ANDREWS, T. S. and HEMBERG, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20** 273–282. https://doi.org/10.1038/s41576-018-0088-9

KRUSKAL, W. H. and WALLIS, W. A. (1952). Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.* **47** 583–621.

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, H., ZHU, B., JIANG, X., GUO, L., XIE, Y., XU, L. and LI, Q. (2024). An interpretable Bayesian clustering approach with feature selection for analyzing spatially resolved transcriptomics data. *Biometrics* **80** ujae066.

LI, J., WITTEN, D. M., JOHNSTONE, I. M. and TIBSHIRANI, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* **13** 523–538.

LI, Q., GUINDANI, M., REICH, B. J., BONDELL, H. D. and VANNUCCI, M. (2017). A Bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints. *Stat. Anal. Data Min.* **10** 393–409. MR3733613 https://doi.org/10.1002/sam.11350

LI, Q., JIANG, S., KOH, A. Y., XIAO, G. and ZHAN, X. (2019). Bayesian modeling of microbiome data for differential abundance analysis. Preprint. Available at arXiv:1902.08741.

LI, Q., ZHANG, M., XIE, Y. and XIAO, G. (2021). Bayesian modeling of spatial molecular profiling data via Gaussian process. *Bioinformatics* **37** 4129–4136.

LIAO, J., LU, X., SHAO, X., ZHU, L. and FAN, X. (2021). Uncovering an organ's molecular architecture at single-cell resolution by spatially resolved transcriptomics. *Trends Biotechnol*. **39** 43–58.

LIU, W., LIAO, X., YANG, Y., LIN, H., YEONG, J., ZHOU, X., SHI, X. and LIU, J. (2022). Joint dimension reduction and clustering analysis of single-cell RNA-seq and spatial transcriptomics data. *Nucleic Acids Res*. **50** e72–e72.

LUBECK, E., COSKUN, A. F., ZHIYENTAYEV, T., AHMAD, M. and CAI, L. (2014). Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11** 360–361. https://doi.org/10.1038/nmeth.2892

LUN, A. T. L., BACH, K. and MARIONI, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol*. **17** 1–14.

MA, S. and HUANG, J. (2017). A concave pairwise fusion approach to subgroup analysis. *J. Amer. Statist. Assoc*. **112** 410–423. MR3646581 https://doi.org/10.1080/01621459.2016.1148039

MA, Y. and ZHOU, X. (2022). Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat. Biotechnol*. **40** 1349–1359.

MAYNARD, K. R., COLLADO-TORRES, L., WEBER, L. M., UYTINGCO, C., BARRY, B. K., WILLIAMS, S. R., CATALLINI, J. L., TRAN, M. N., BESICH, Z. et al. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci*. **24** 425–436.

MILLER, J. W. and HARRISON, M. T. (2018). Mixture models with a prior on the number of components. *J. Amer. Statist. Assoc*. **113** 340–356. MR3803469 https://doi.org/10.1080/01621459.2016.1255636

MOSES, L. and PACHTER, L. (2022). Museum of spatial transcriptomics. *Nat. Methods* **19** 534–546.

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

NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176. https://doi.org/10.1093/biostatistics/5.2.155

PAN, W. and SHEN, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res*. **8** 1145–1164.

PAPALEXI, E. and SATIJA, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev., Immunol*. **18** 35–45. https://doi.org/10.1038/nri.2017.76

PETERSON, C., STINGO, F. C. and VANNUCCI, M. (2015). Bayesian inference of multiple Gaussian graphical models. *J. Amer. Statist. Assoc*. **110** 159–174. MR3338494 https://doi.org/10.1080/01621459.2014.896806

PHAM, D., TAN, X., XU, J., GRICE, L. F., LAM, P. Y., RAGHUBAR, A., VUKOVIC, J., RUITENBERG, M. J. and NGUYEN, Q. (2020). stLearn: Integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *BioRxiv*. https://doi.org/10.1101/2020.05.31.125658.

RISSO, D., PERRAUDEAU, F., GRIBKOVA, S., DUDOIT, S. and VERT, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun*. **9** 284.

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.

SANTOS, J. M. and EMBRECHTS, M. (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification. In *International Conference on Artificial Neural Networks* 175–184. Springer, Berlin.

SATIJA, R., FARRELL, J. A., GENNERT, D., SCHIER, A. F. and REGEV, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol*. **33** 495–502. https://doi.org/10.1038/nbt.3192

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

SONG, F., CHAN, G. M. A. and WEI, Y. (2020). Flexible experimental designs for valid single-cell RNA-sequencing experiments allowing batch effects correction. *Nat. Commun*. **11** 3274.

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.

STINGO, F. C., GUINDANI, M., VANNUCCI, M. and CALHOUN, V. D. (2013). An integrative Bayesian modeling approach to imaging genetics. *J. Amer. Statist. Assoc*. **108** 876–891. MR3174670 https://doi.org/10.1080/01621459.2013.804409

SUN, S., ZHU, J. and ZHOU, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17** 193–200.

SVENSSON, V., TEICHMANN, S. A. and STEGLE, O. (2018). SpatialDE: Identification of spatially variable genes. *Nat. Methods* **15** 343–346.

TADESSE, M. G., SHA, N. and VANNUCCI, M. (2005). Bayesian variable selection in clustering high-dimensional data. *J. Amer. Statist. Assoc.* **100** 602–617. MR2160563 https://doi.org/10.1198/016214504000001565

TANAY, A. and REGEV, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541** 331–338. https://doi.org/10.1038/nature21350

THRANE, K., ERIKSSON, H., MAASKOLA, J., HANSSON, J. and LUNDEBERG, J. (2018). Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res.* **78** 5970–5979. https://doi.org/10.1158/0008-5472.CAN-18-0747

VAN DER MAATEN, L. and HINTON, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** 2579–2605.

VICKOVIC, S., ERASLAN, G., SALMÉN, F., KLUGHAMMER, J., STENBECK, L., SCHAPIRO, D., ÄIJÖ, T., BONNEAU, R., BERGENSTRÅHLE, L. et al. (2019). High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods* **16** 987–990.

WANG, X., ALLEN, W. E., WRIGHT, M. A., SYLWESTRAK, E. L., SAMUSIK, N., VESUNA, S., EVANS, K., LIU, C., RAMAKRISHNAN, C. et al. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361** eaat5691.

WITTEN, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *Ann. Appl. Stat.* **5** 2493–2518. MR2907124 https://doi.org/10.1214/11-AOAS493

WU, Q. and LUO, X. (2022). Estimating heterogeneous gene regulatory networks from zero-inflated single-cell expression data. *Ann. Appl. Stat.* **16** 2183–2200. MR4489205 https://doi.org/10.1214/21-aoas1582

YAN, Y. and LUO, X. (2024). Bayesian integrative region segmentation in spatially resolved transcriptomic studies. *J. Amer. Statist. Assoc.* **119** 1709–1721. MR4797896 https://doi.org/10.1080/01621459.2024.2308323

YANG, Y., SHI, X., LIU, W., ZHOU, Q., CHAN LAU, M., CHUN TATT LIM, J., SUN, L., NG, C. C. Y., YEONG, J. et al. (2022). SC-MEB: Spatial clustering with hidden Markov random field using empirical Bayes. *Brief. Bioinform.* **23** bbab466.

ZHAO, E., STONE, M. R., REN, X., GUENTHOER, J., SMYTHE, K. S., PULLIAM, T., WILLIAMS, S. R., UYTINGCO, C. R., TAYLOR, S. E. et al. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.* **39** 1375–1384.

ZHAO, P., YANG, H.-C., DEY, D. K. and HU, G. (2023). Spatial clustering regression of count value data via Bayesian mixture of finite mixtures. In *Proceedings of the* 29*th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 3504–3512.

ZHU, B., HU, G., XU, L., FAN, X. and LI, Q. (2025). Supplement to "Bayesian nonparametric clustering with feature selection for spatially resolved transcriptomics data." https://doi.org/10.1214/25-AOAS2014SUPPA, https://doi.org/10.1214/25-AOAS2014SUPPB

ZHU, Q., SHAH, S., DRIES, R., CAI, L. and YUAN, G.-C. (2018). Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nat. Biotechnol.* **36** 1183–1190.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1048–1064
https://doi.org/10.1214/25-AOAS2029

COMPUTATIONALLY EFFICIENT WHOLE-GENOME SIGNAL REGION DETECTION FOR QUANTITATIVE AND BINARY TRAITS

BY FAN WANG[1,a], WEI ZHANG[2,b] AND FANG YAO[2,c]

[1]*Department of Biostatistics, Mailman School of Public Health, Columbia University,* [a]*fw2400@cumc.columbia.edu*
[2]*School of Mathematical Sciences, Center for Statistical Science, Peking University,* [b]*wei_zhang@stu.pku.edu.cn,* [c]*fyao@math.pku.edu.cn*

The identification of genetic signal regions in the human genome is critical for understanding the genetic architecture of complex traits and diseases. Numerous methods based on scan algorithms (i.e., QSCAN, SCANG, SCANG-STAAR) have been developed to allow dynamic window sizes in whole-genome association studies. Beyond scan algorithms, we have recently developed the binary and research (BiRS) algorithm, which is more computationally efficient than scan-based methods and exhibits superior statistical power. However, the BiRS algorithm is based on two-sample mean test for binary traits, not accounting for multidimensional covariates or nonbinary outcomes. In this work we propose a new maximal score test based on summary statistics computed from a generalized linear model, which accommodates regression-based statistics and allows testing of both continuous and binary outcomes. We then present a distributed version of the BiRS algorithm (dBiRS) that incorporates this new test, enabling parallel computing of blockwise results by aggregation through a central machine to ensure both detection accuracy and computational efficiency, which has theoretical guarantees for controlling familywise error rates and false discovery rates while maintaining the power advantages of the original algorithm. Applying dBiRS to detect genetic regions associated with fluid intelligence and prospective memory using whole-exome sequencing data from the UK Biobank, we validate previous findings and identify numerous novel rare variants near newly implicated genes. These discoveries offer valuable insights into the genetic basis of cognitive performance and neurodegenerative disorders, highlighting the potential of dBiRS as a scalable and powerful tool for whole-genome signal region detection.

REFERENCES

BAKSHI, A., ZHU, Z., VINKHUYZEN, A. A. et al. (2016). Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Sci*. *Rep*. **6** 1–9.

CAI, T. T. and WEI, H. (2022). Distributed adaptive Gaussian mean estimation with unknown variance: Interactive protocol helps adaptation. *Ann*. *Statist*. **50** 1992–2020. MR4474480 https://doi.org/10.1214/21-aos2167

DAVIES, G., LAM, M., HARRIS, S. E., TRAMPUSH, J. W., LUCIANO, M., HILL, W. D., HAGENAARS, S. P., RITCHIE, S. J., MARIONI, R. E. et al. (2018). Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat*. *Commun*. **9**. 2098.

HE, Z., LE GUEN, Y., LIU, L. et al. (2021a). Genome-wide analysis of common and rare variants via multiple knockoffs at biobank scale, with an application to Alzheimer disease genetics. *Am*. *J*. *Hum*. *Genet*. **108** 2336–2353.

HE, Z., LIU, L., WANG, C. et al. (2021b). Identification of putative causal loci in whole-genome sequencing data via knockoff statistics. *Nat*. *Commun*. **12** 3152.

HEROLD, C., HOOLI, B. V., MULLIN, K., LIU, T., ROEHR, J. T., MATTHEISEN, M., PARRADO, A. R., BERTRAM, L., LANGE, C. et al. (2016). Family-based association analyses of imputed genotypes reveal genome-wide significant association of Alzheimer's disease with OSBPL6, PTPRG, and PDCL3. *Mol*. *Psychiatry* **21** 1608–1612. https://doi.org/10.1038/mp.2015.218

Key words and phrases. Familywise error rate, signal detection, distributed learning, whole genome association studies.

HOLLINS, S. L. and CAIRNS, M. J. (2016). MicroRNA: Small RNA mediators of the brains genomic response to environmental stress. *Prog. Neurobiol.* **143** 61–81. https://doi.org/10.1016/j.pneurobio.2016.06.005

JENG, X. J., CAI, T. T. and LI, H. (2010). Optimal sparse segment identification with application in copy number variation analysis. *J. Amer. Statist. Assoc.* **105** 1156–1166. MR2752611 https://doi.org/10.1198/jasa.2010.tm10083

KANKI, H., SASAKI, T., MATSUMURA, S., KAWANO, T., TODO, K., OKAZAKI, S., NISHIYAMA, K., TAKEMORI, H. and MOCHIZUKI, H. (2020). CREB coactivator CRTC2 plays a crucial role in endothelial function. *J. Neurosci.* **40** 9533–9546.

KING, I. F., YANDAVA, C. N., MABB, A. M., HSIAO, J. S., HUANG, H.-S., PEARSON, B. L., CALABRESE, J. M., STARMER, J., PARKER, J. S. et al. (2013). Topoisomerases facilitate transcription of long genes linked to autism. *Nature* **501** 58–62.

KIRCHER, M., WITTEN, D. M., JAIN, P., O'ROAK, B. J., COOPER, G. M. and SHENDURE, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46** 310–315. https://doi.org/10.1038/ng.2892

LEE, J. J., WEDOW, R., OKBAY, A., KONG, E., MAGHZIAN, O., ZACHER, M., NGUYEN-VIET, T. A., BOWERS, P., SIDORENKO, J. et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50** 1112–1121.

LI, X., LI, Z., ZHOU, H. et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52** 969–983.

LI, Z., LI, X., LIU, Y., SHEN, J., CHEN, H., ZHOU, H., MORRISON, A. C., BOERWINKLE, E. and LIN, X. (2019). Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. *Am. J. Hum. Genet.* **104** 802–814.

LI, Z., LI, X., ZHOU, H. et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nat. Methods* **19** 1599–1611.

LI, Z., LIU, Y. and LIN, X. (2020). Simultaneous detection of signal regions using quadratic scan statistics with applications to whole genome association studies. *J. Amer. Statist. Assoc.* **117** 823–834. MR4436315 https://doi.org/10.1080/01621459.2020.1822849

LI, Z., LIU, Y. and LIN, X. (2022). Simultaneous detection of signal regions using quadratic scan statistics with applications to whole genome association studies. *J. Amer. Statist. Assoc.* **117** 823–834. MR4436315 https://doi.org/10.1080/01621459.2020.1822849

LIU, J. Z., MCRAE, A. F., NYHOLT, D. R. et al. (2010). A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87** 139–145.

MADABHUSHI, R., GAO, F., PFENNING, A. R., PAN, L., YAMAKAWA, S., SEO, J., RUEDA, R., PHAN, T. X., YAMAKAWA, H. et al. (2015). Activity-induced DNA breaks govern the expression of neuronal early-response genes. *Cell* **161** 1592–1605.

MADSEN, B. E. and BROWNING, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5** 1–11.

MANDAL, M., MAIENSCHEIN-CLINE, M., MAFFUCCI, P., VESELITS, M., KENNEDY, D. E., MCLEAN, K. C., OKOREEH, M. K., KARKI, S., CUNNINGHAM-RUNDLES, C. et al. (2018). BRWD1 orchestrates epigenetic landscape of late B lymphopoiesis. *Nat. Commun.* **9** 3888. https://doi.org/10.1038/s41467-018-06165-6

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R. et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.

MEIER, S. M., TRONTTI, K., PURVES, K. L., ALS, T. D., GROVE, J., LAINE, M., PEDERSEN, M. G., BYBJERG-GRAUHOLM, J., BÆKVED-HANSEN, M. et al. (2019). Genetic variants associated with anxiety and stress-related disorders: A genome-wide association study and mouse-model study. *JAMA Psychiatr.* **76** 924–932.

MORGENTHALER, S. and THILLY, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutat. Res.* **615** 28–56. https://doi.org/10.1016/j.mrfmmm.2006.09.003

NAUS, J. I. (1982). Approximations for distributions of scan statistics. *J. Amer. Statist. Assoc.* **77** 177–183. MR0648042

OKBAY, A., WU, Y., WANG, N., JAYASHANKAR, H., BENNETT, M., NEHZATI, S. M., SIDORENKO, J., KWEON, H., GOLDMAN, G. et al. (2022). Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat. Genet.* **54** 437–449.

OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572. https://doi.org/10.1093/biostatistics/kxh008

PASMAN, J. A., DEMANGE, P. A., GULOKSUZ, S., WILLEMSEN, A., ABDELLAOUI, A., TEN HAVE, M., HOTTENGA, J.-J., BOOMSMA, D. I., DE GEUS, E. et al. (2022). Genetic risk for smoking: Disentangling interplay between genes and socioeconomic status. *Behav. Genet.* **52** 92–107.

RIPKE, S., SANDERS, A., KENDLER, K., LEVINSON, D., SKLAR, P., HOLMANS, P., LIN, D., DUAN, J., OPHOFF, R. et al. (2011). Schizophrenia psychiatric genome-wide association study (GWAS) consortium genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43** 969–976.

SABETI, I. S. P. C. and SCHAFFNER, S. F. (2014). Cosi2: An efficient simulator of exact and approximate coalescent with selection. *Bioinformatics* **30** 3427–3429.

SAVAGE, J. E., JANSEN, P. R., STRINGER, S., WATANABE, K., BRYOIS, J., DE LEEUW, C. A., NAGEL, M., AWASTHI, S., BARR, P. B. et al. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50** 912–919.

SOLLIS, E., MOSAKU, A., ABID, A., BUNIELLO, A., CEREZO, M., GIL, L., GROZA, T., GÜNEŞ, O., HALL, P. et al. (2023). The NHGRI-EBI GWAS catalog: Knowledgebase and deposition resource. *Nucleic Acids Res.* **51** D977–D985.

TIWARI, V. K., BURGER, L., NIKOLETOPOULOU, V., DEOGRACIAS, R., THAKURELA, S., WIRBELAUER, C., KAUT, J., TERRANOVA, R., HOERNER, L. et al. (2012). Target genes of topoisomerase IIβ regulate neuronal survival and are defined by their chromatin state. *Proc. Natl. Acad. Sci. USA* **109** E934–E943.

VISSCHER, P. M., WRAY, N. R., ZHANG, Q. et al. (2017). 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **101** 5–22.

WANG, F., ZHANG, W. and YAO, F. (2025). Supplement to “Computationally Efficient Whole-Genome Signal Region Detection for Quantitative and Binary Traits.” https://doi.org/10.1214/25-AOAS2029SUPP

WANG, H., YANG, J., SCHNEIDER, J. A., DE JAGER, P. L., BENNETT, D. A. and ZHANG, H.-Y. (2020). Genome-wide interaction analysis of pathological hallmarks in Alzheimer’s disease. *Neurobiol. Aging* **93** 61–68.

WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89** 82–93.

XUE, K. and YAO, F. (2020). Distribution and correlation-free two-sample test of high-dimensional means. *Ann. Statist.* **48** 1304–1328. MR4124324 https://doi.org/10.1214/19-AOS1848

ZHANG, L., PAN, Y., HUANG, G., LIANG, Z., LI, L., ZHANG, M. and ZHANG, Z. (2023). A brain-wide genome-wide association study of candidate quantitative trait loci associated with structural and functional phenotypes of pain sensitivity. *Cereb. Cortex* **33** 7297–7309.

ZHANG, N. R., SIEGMUND, D. O., JI, H. and LI, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97** 631–645. MR2672488 https://doi.org/10.1093/biomet/asq025

ZHANG, W., WANG, F. and YAO, F. (2025). Fast signal region detection with application to whole genome association studies. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2025.2464271

ZHOU, H., ARAPOGLOU, T., LI, X. et al. (2023). FAVOR: Functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Res.* **51** D1300–D1311.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1065–1085
https://doi.org/10.1214/25-AOAS2016

IS THE PRICE CAP FOR GAS USEFUL? EVIDENCE FROM EUROPEAN COUNTRIES

BY FRANCESCO RAVAZZOLO[1,2,a] AND LUCA ROSSINI[3,4,b]

[1]*Department of Data Science and Analytics, BI Norwegian Business School*
[2]*Faculty of Economics and Management, Free University of Bozen-Bolzano,* [a]*francesco.ravazzolo@unibz.it*
[3]*Department of Economics, Management and Quantitative Methods, University of Milan*
[4]*Fondazione Eni Enrico Mattei,* [b]*luca.rossini@unimi.it*

Since Russia's invasion of Ukraine, many countries have pledged to end or restrict their oil and gas imports to curtail Moscow's revenues and hinder its war effort. Thus, the European ministers agreed to trigger a cap on the gas price. To detect the importance of the price cap for gas, we provide a mixture representation for the gas price to detect the presence of outliers made by a truncated normal distribution and a uniform one. We focus our analysis on a unique dataset of different commodity prices for Germany and Italy, which are major Russian gas importers by exploiting the response of the different commodities to a gas shock through a Bayesian vector autoregressive (VAR) model. As a result, including a lower gas price cap smooths the impact of a gas shock on electricity prices, while not considering a price cap will increase exponentially this impact. Regarding the other commodities, gas shocks matter in the short and long run when a price cap is not considered.

REFERENCES

AASTVEIT, K. A., BJØRNLAND, H. C. and CROSS, J. L. (2023). Inflation expectations and the pass-through of oil prices. *Rev. Econ. Stat.* **105** 733–743. https://doi.org/10.1162/rest_a_01073

ADOLFSEN, J. F., FERRARI MINESSO, M., MORK, J. E. and VAN ROBAYS, I. (2024). Gas price shocks and euro area inflation. *J. Int. Money Financ.* **149** 103183. https://doi.org/10.1016/j.jimonfin.2024.103183

AMIRNEKOOEI, K., ARDEHALI, M. M. and SADRI, A. (2017). Optimal energy pricing for integrated natural gas and electric power network with considerations for techno-economic constraints. *Energy* **123** 693–709. https://doi.org/10.1016/j.energy.2017.01.145

BAUMEISTER, C. and HAMILTON, J. D. (2019). Structural interpretation of vector autoregressions with incomplete identification: Revisiting the role of oil supply and demand shocks. *Amer. Econ. Rev.* **109** 1873–1910. https://doi.org/10.1257/aer.20151569

BAUMEISTER, C., KOROBILIS, D. and LEE, T. K. (2022). Energy markets and global economic conditions. *Rev. Econ. Stat.* **104** 828–844. https://doi.org/10.1162/rest_a_00977

BJØRNLAND, H. C., CROSS, J. L. and KAPFHAMMER, F. (2023). The drivers of emission reductions in the European carbon market. Technical Report No. 08. CAMP Working Paper Series.

BJØRNLAND, H. C. and THORSRUD, L. A. (2019). Commodity prices and fiscal policy design: Procyclical despite a rule. *J. Appl. Econometrics* **34** 161–180. MR3922625 https://doi.org/10.1002/jae.2669

CARRIERO, A., CHAN, J., CLARK, T. E. and MARCELLINO, M. (2022). Corrigendum to "Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors" [J. Econometrics 212 (1) (2019) 137–154]. *J. Econometrics* **227** 506–512. MR4384683 https://doi.org/10.1016/j.jeconom.2021.11.010

CASOLI, C., MANERA, M. and VALENTI, D. (2024). Energy shocks in the Euro area: Disentangling the pass-through from oil and gas prices to inflation. *J. Int. Money Financ.* **147** 103154. https://doi.org/10.1016/j.jimonfin.2024.103154

CHULIÁ, H., FURIÓ, D. and URIBE, J. M. (2019). Volatility spillovers in energy markets. *Energy J.* **40** 173–198.

EUROPEAN COMMISSION (2022). European commission energy news.

DIAGOUPIS, T. D., ANDRIANESIS, P. E. and DIALYNAS, E. N. (2016). A planning approach for reducing the impact of natural gas network on electricity markets. *Appl. Energy* **175** 189–198. https://doi.org/10.1016/j.apenergy.2016.05.006

Key words and phrases. Bayesian time series, forecasted error variance decomposition, gas price cap, impulse response function, mixture representation.

ECONOMOU, T., JOHNSON, C. and DYSON, E. (2024). A hierarchical spline model for correcting and hindcasting temperature data. *Ann. Appl. Stat.* **18** 1709–1728. MR4728686 https://doi.org/10.1214/23-aoas1855

EHRART, K.-M. and SCHLECHT, I. (2022). Introducing a price cap on Russian gas: A game theoretic analysis. Technical report, ZBW—Leibniz Information Centre for Economics.

EHRHART, K.-M., SCHLECHT, I., SCHMITZ, J. and WANG, R. (2023). Comparison of price caps and tariffs to counter a foreign monopoly. *Econom. Lett.* **227** Paper No. 111128, 4 pp. MR4584940 https://doi.org/10.1016/j.econlet.2023.111128

FEZZI, C. and BUNN, D. (2010). Structural analysis of electricity demand and supply interactions. *Oxf. Bull. Econ. Stat.* **72** 827–856. https://doi.org/10.1111/j.1468-0084.2010.00596.x

GÜNTNER, J., REIF, M. and WOLTERS, M. (2024). Sudden stop: Supply and demand shocks in the German natural gas market. *J. Appl. Econometrics* **39** 1282–1300. MR4835338 https://doi.org/10.1002/jae.3089

HIRTH, L. (2018). What caused the drop in European electricity prices? A factor decomposition analysis. *Energy J.* **39** 143–158.

KILIAN, L. and ZHOU, X. (2022). Oil prices, gasoline prices, and inflation expectations. *J. Appl. Econometrics* **37** 867–881. MR4470807 https://doi.org/10.1002/jae.2911

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

KOROBILIS, D. and SHIMIZU, K. (2022). Bayesian approaches to shrinkage and sparse estimation. *Found Trends Econom.* **11** 230–354. https://doi.org/10.1561/0800000041

KYRITSIS, E. and ANDERSSON, J. (2019). Causality in quantiles and dynamic relations in energy markets: (De)tails matter. *Energy Policy* **133** 110933.

MARTÍNEZ, B. and TORRÓ, H. (2018). Hedging spark spread risk with futures. *Energy Policy* **113** 731–746. https://doi.org/10.1016/j.enpol.2017.11.038

MILLS, A., WISER, R., MILLSTEIN, D., CARVALLO, J. P., GORMAN, W., SEEL, J. and JEONG, S. (2021). The impact of wind, solar, and other factors on the decline in wholesale power prices in the United States. *Appl. Energy* **283** 116266. https://doi.org/10.1016/j.apenergy.2020.116266

NAKAJIMA, T. and HAMORI, S. (2013). Testing causal relationships between wholesale electricity prices and primary energy prices. *Energy Policy* **62** 869–877. https://doi.org/10.1016/j.enpol.2013.07.033

UNITED NATIONS (2022). COP26.

OHLER, A., MOHAMMADI, H. and LOOMIS, D. G. (2020). Electricity restructuring and the relationship between fuel costs and electricity prices for industrial and residential customers. *Energy Policy* **142** 111559. https://doi.org/10.1016/j.enpol.2020.111559

RAVAZZOLO, F. and ROSSINI, L. (2025). Supplement to "Is the price cap for Gas useful? Evidence from European countries." https://doi.org/10.1214/25-AOAS2016SUPPA, https://doi.org/10.1214/25-AOAS2016SUPPB

ROEGER, W. and WELFENS, P. J. J. (2022). Gas price caps and electricity production effects in the context of the Russo-Ukrainian war: Modeling and new policy reforms. *Int. Econ. Econ. Policy* **19** 645–673. https://doi.org/10.1007/s10368-022-00552-7

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

SCARCIOFFOLO, A. R. and ETIENNE, X. (2021). Testing directional predictability between energy prices: A quantile-based analysis. *Resour. Policy* **74** 102258.

URIBE, J. M., GUILLEN, M. and MOSQUERA-LÓPEZ, S. (2018). Uncovering the nonlinear predictive causality between natural gas and electricity prices. *Energy Econ.* **74** 904–916. https://doi.org/10.1016/j.eneco.2018.07.025

URIBE, J. M., MOSQUERA-LÓPEZ, S. and ARENAS, O. J. (2022). Assessing the relationship between electricity and natural gas prices in European markets in times of distress. *Energy Policy* **166** 113018. https://doi.org/10.1016/j.enpol.2022.113018

WOO, C.-K., OLSON, A., HOROWITZ, I. and LUK, S. (2006). Bi-directional causality in California's electricity and natural-gas markets. *Energy Policy* **34** 2060–2070. https://doi.org/10.1016/j.enpol.2005.02.016

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1086–1109
https://doi.org/10.1214/24-AOAS1923

ANALYZING ENVIRONMENTAL BIOASSAYS WITH SPATIAL ODDS, RISK, AND SURVIVAL PROBABILITY RATIO REGRESSIONS

BY DEBASHIS MONDAL[1,a] AND XIAOHUI CHANG[2,b]

[1]*Department of Statistics and Data Science, Washington University in St. Louis,* [a]*mondal@wustl.edu*
[2]*College of Business, Oregon State University,* [b]*xiaohui.chang@oregonstate.edu*

Environmental bioassays, such as sediment toxicity tests, provide a broad survey of toxicity that is crucial for the conservation and protection of marine and estuarine ecosystems. Using odds, risk, and survival probability ratios, this paper presents a critical evaluation of sediment toxicity tests data collected in the New York–New Jersey harbor area. It further derives spatial regression analysis to combine test results, predict toxicity at unsampled locations, and determine the effects of specific contaminants. The proposed spatial analysis is based on non-Euclidean distances and is applicable to complex sampling domains with nonconvex boundaries. The findings suggest that current practices can be improved with the use of relevant statistical methods with odds, risk, survival probability ratios, and attributable effects.

REFERENCES

ABBOTT, W. S. (1925). A method of computing the effectiveness of an insecticide. *J. Econ. Entomol.* **18** 265–267.

ADAMS, D. and BENYI, S. (2003). Sediment quality of the NY/NJ Harbor system: A 5-year revisit. An investigation under the Regional Environmental Monitoring and Assessment Program (REMAP). Technical Report, US Environmental Protection Agency—Region 2, Edison, NJ.

AGRESTI, A. (1999). On logit confidence intervals for the odds ratio with small samples. *Biometrics* **55** 597–602. https://doi.org/10.1111/j.0006-341x.1999.00597.x

AGRESTI, A. and MIN, Y. (2002). Unconditional small-sample confidence intervals for the odds ratio. *Biostatistics* **3** 379–386. https://doi.org/10.1093/biostatistics/3.3.379

ARU, J., LUPU, T. and SEPÚLVEDA, A. (2020). The first passage sets of the 2D Gaussian free field: Convergence and isomorphisms. *Comm. Math. Phys.* **375** 1885–1929. MR4091511 https://doi.org/10.1007/s00220-020-03718-z

BAPAT, R. B. (2004). Resistance matrix of a weighted graph. *MATCH Commun. Math. Comput. Chem.* **50** 73–82. MR2037425

BESAG, J. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746. MR1380811 https://doi.org/10.1093/biomet/82.4.733

BESAG, J. and MONDAL, D. (2005). First-order intrinsic autoregressions and the de Wijs process. *Biometrika* **92** 909–920. MR2234194 https://doi.org/10.1093/biomet/92.4.909

BESAG, J., YORK, J. and MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43** 1–59. MR1105822 https://doi.org/10.1007/BF00116466

CHAPMAN, P. M. (1990). The sediment quality triad approach to determining pollution-induced degradation. *Sci. Total Environ.* **97–98** 815–825.

CHATTERJEE, S., LAHIRI, P. and LI, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Ann. Statist.* **36** 1221–1245. MR2418655 https://doi.org/10.1214/07-AOS512

CLEUVERS, M. (2003). Aquatic ecotoxicity of pharmaceuticals including the assessment of combination effects. *Toxicol. Lett.* **142** 185–194. https://doi.org/10.1016/s0378-4274(03)00068-7

CURRIERO, F. C. (2006). On the use of non-Euclidean distance measures in geostatistics. *Math. Geol.* **38** 907–926. MR2325912 https://doi.org/10.1007/s11004-006-9055-7

DARROCH, J. and BORKENT, M. (1994). Synergism, attributable risk and interaction for two binary exposure factors. *Biometrika* **81** 259–270.

DAVIS, C. D., EPPS, C. W., FLITCROFT, R. L. and BANKS, M. A. (2018). Refining and defining riverscape genetics: How rivers influence population genetic structure. *Wiley Interdiscip. Rev. Water* **5** e1269.

Key words and phrases. Sediment toxicity test, 2×2 tables, survival probability ratio, attributable effect, spatial bootstrap, intrinsic autoregression, de Wijs process, resistance distance, h-likelihood, REML.

DICKSON, B. G., ALBANO, C. M., ANANTHARAMAN, R., BEIER, P., FARGIONE, J., GRAVES, T. A., GRAY, M. E., HALL, K. R., LAWLER, J. J. et al. (2019). Circuit-theory applications to connectivity science and conservation. *Conserv. Biol.* **33** 239–249.

DIGGLE, P. J., MENEZES, R. and SU, T. (2010). Geostatistical inference under preferential sampling. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59** 191–232. MR2744471 https://doi.org/10.1111/j.1467-9876.2009.00701.x

DI TORO, D. M., MAHONY, J. D., HANSEN, D. J., SCOTT, K. J., HICKS, M. B., MAYR, S. M. and REDMOND, M. S. (1990). Toxicity of cadmium in sediments: The role of acid volatile sulfide. *Environ. Toxicol. Chem.* **9** 1487–1502.

DOYLE, P. G. and SNELL, J. L. (1984). *Random Walks and Electric Networks. Carus Mathematical Monographs* **22**. Mathematical Association of America, Washington, DC. MR0920811

DUDGEON, D., ARTHINGTON, A. H., GESSNER, M. O., KAWABATA, Z.-I., KNOWLER, D. J., LÉVÊQUE, C., NAIMAN, R. J., PRIEUR-RICHARD, A.-H., SOTO, D. et al. (2006). Freshwater biodiversity: Importance, threats, status and conservation challenges. *Biol. Rev.* **81** 163–182.

DUFFIN, R. J. and SHAFFER, D. H. (1960). Asymptotic expansion of double Fourier transforms. *Duke Math. J.* **27** 581–596. MR0117501

DUTTA, S. and MONDAL, D. (2015). An h-likelihood method for spatial mixed linear models based on intrinsic auto-regressions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 699–726. MR3351451 https://doi.org/10.1111/rssb.12084

DUTTA, S. and MONDAL, D. (2016). REML estimation with intrinsic Matérn dependence in the spatial linear mixed model. *Electron. J. Stat.* **10** 2856–2893. MR3553914 https://doi.org/10.1214/16-EJS1125

EPA (1993). Environmental Monitoring and Assessment Program (EMAP) laboratory methods manual estuaries. Technical Report, U.S. Environmental Protection Agency, Office of Research and Development, Environmental Monitoring Systems Laboratory Cincinnati, OH.

FAY, M. P. and HUNSBERGER, S. A. (2021). Practical valid inferences for the two-sample binomial problem. *Stat. Surv.* **15** 72–110. MR4255287 https://doi.org/10.1214/21-ss131

FISHER, R. A. (1934). *Statistical Methods for Research Workers*. Oliver and Bovdl, Edinburgh.

FUSHIKI, T., KOMAKI, F. and AIHARA, K. (2004). On parametric bootstrapping and Bayesian prediction. *Scand. J. Stat.* **31** 403–416. MR2087833 https://doi.org/10.1111/j.1467-9469.2004.02_127.x

GART, J. J. and NAM, J. (1988). Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics* **44** 323–338. MR0945766 https://doi.org/10.2307/2531848

HÄDER, D.-P. and ERZINGER, G. S. (2017). Daphniatox—online monitoring of aquatic pollution and toxic substances. *Chemosphere* **167** 228–235. https://doi.org/10.1016/j.chemosphere.2016.09.155

HAMILTON, M. A. (1979). Choosing the parameter for a 2×2 table or a $2 \times 2 \times 2$ table analysis. *Amer. J. Epidemiol.* **109** 362–375. https://doi.org/10.1093/oxfordjournals.aje.a112689

HARVILLE, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72** 320–340. MR0451550

KERN, J. C. and HIGDON, D. M. (1999). A distance metric to account for edge effects in spatial data analysis. In *Proceedings-American Statistical Association Biometrics Section*.

KLEIN, D. J. and RANDIC, M. (1993). Resistance distance. *J. Math. Chem.* **12** 81–95.

LEE, Y. and NELDER, J. A. (1996). Hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser. B* **58** 619–678. MR1410182

LEE, Y. and NELDER, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* **88** 987–1006. MR1872215 https://doi.org/10.1093/biomet/88.4.987

LIANG, K.-Y. (1983). On information and ancillarity in the presence of a nuisance parameter. *Biometrika* **70** 607–612. MR0725374 https://doi.org/10.1093/biomet/70.3.607

LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. MR2853727 https://doi.org/10.1111/j.1467-9868.2011.00777.x

LITTLE, L. S., EDWARDS, D. and PORTER, D. E. (1997). Kriging in estuaries: As the crow flies, or as the fish swims? *J. Exp. Mar. Biol. Ecol.* **213** 1–11.

MATHERON, G. (1971). The theory of regionalised variables and its applications. *Les Cahiers Du Centre de Morphologie Mathematique* **5** 212.

MCCULLAGH, P. (2002). What is a statistical model? *Ann. Statist.* **30** 1225–1310. MR1936320 https://doi.org/10.1214/aos/1035844977

MCCULLAGH, P. and CLIFFORD, D. (2006). Evidence for conformal invariance of crop yields. *Proc. R. Soc. A, Math. Phys. Eng. Sci.* **462** 2119–2143.

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

MEHTA, C. R., PATEL, N. R. and GRAY, R. (1985). Computing an exact confidence interval for the common odds ratio in several 2×2 contingency tables. *J. Amer. Statist. Assoc.* **80** 969–973. MR0819602

MONDAL, D. (2005). Variogram calculations for first-order intrinsic autoregressions and the de Wijs process. Technical Report No. 479, Department of Statistics, Univ. Washington, Seattle.

MONDAL, D. (2015). Applying Dynkin's isomorphism: An alternative approach to understand the Markov property of the de Wijs process. *Bernoulli* **21** 1289–1303. MR3352044 https://doi.org/10.3150/13-BEJ541

MONDAL, D. (2018). On edge correction of conditional and intrinsic autoregressions. *Biometrika* **105** 447–454. MR3804413 https://doi.org/10.1093/biomet/asy014

MONDAL, D. (2019). Generalized Gaussian Markov random fields and modeling disease risk. Under revision.

MONDAL, D. and CHANG, X. (2025). Supplement to "Analyzing Environmental Bioassays with Spatial Odds, Risk, and Survival Probability Ratio Regressions." https://doi.org/10.1214/24-AOAS1923SUPP

OGENDI, G. M., BRUMBAUGH, W. G., HANNIGAN, R. E. and FARRIS, J. L. (2007). Effects of acid-volatile sulfide on metal bioavailability and toxicity to midge (Chironomus tentans) larvae in black shale sediments. *Environ. Toxicol. Chem.* **26** 325–334. https://doi.org/10.1897/06-080r.1

PETTIGREW, H. M., GART, J. J. and THOMAS, D. G. (1986). The bias and higher cumulants of the logarithm of a binomial variate. *Biometrika* **73** 425–435. MR0855902 https://doi.org/10.1093/biomet/73.2.425

PIEGORSCH, W. and BAILER, A. J. (1997). *Statistics for Environmental Biology and Toxicology*. CRC Press, London.

PLACKETT, R. L. and HEWLETT, P. S. (1952a). Quantal responses to mixtures of poisons. *J. Roy. Statist. Soc. Ser. B* **14** 141–154.

PLACKETT, R. L. and HEWLETT, P. S. (1952b). Similar joint action of insecticides. *Nature* **169** 198–199.

RATHBUN, S. L. (1998). Spatial modelling in irregularly shaped regions: Kriging estuaries. *Environmetrics* **9** 109–129.

RICHARDSON, T. S., ROBINS, J. M. and WANG, L. (2017). On modeling and estimation for the relative risk and risk difference. *J. Amer. Statist. Assoc.* **112** 1121–1130. MR3735364 https://doi.org/10.1080/01621459.2016.1192546

ROBERTSON, J. L., JONES, M. M., OLGUIN, E. and ALBERTS, B. (2017). *Bioassays with Arthropods*. CRC Press, Boca Raton, FL.

ROBINSON, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statist. Sci.* **6** 15–51. MR1108815

ROSENBAUM, P. R. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika* **88** 219–231. MR1841270 https://doi.org/10.1093/biomet/88.1.219

ROSENBAUM, P. R. (2002). Attributing effects to treatment in matched observational studies. *J. Amer. Statist. Assoc.* **97** 183–192. MR1963391 https://doi.org/10.1198/016214502753479329

ROSENBERG, D. M. and RESH, V. H. (1993). *Freshwater Biomonitoring and Benthic Invertebrates*. CRC Press, London and New York.

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

SCHRAMM, O. and SHEFFIELD, S. (2009). Contour lines of the two-dimensional discrete Gaussian free field. *Acta Math.* **202** 21–137. MR2486487 https://doi.org/10.1007/s11511-009-0034-y

SCOTT-HAYWARD, L. A. S., MACKENZIE, M. L., DONOVAN, C. R., WALKER, C. G. and ASHE, E. (2014). Complex region spatial smoother (CReSS). *J. Comput. Graph. Statist.* **23** 340–360. MR3215814 https://doi.org/10.1080/10618600.2012.762920

SHEFFIELD, S. (2007). Gaussian free fields for mathematicians. *Probab. Theory Related Fields* **139** 521–541. MR2322706 https://doi.org/10.1007/s00440-006-0050-1

TETALI, P. (1991). Random walks and the effective resistance of networks. *J. Theoret. Probab.* **4** 101–109. MR1088395 https://doi.org/10.1007/BF01046996

THOMAS, D. G. and GART, J. J. (1977). A table of exact confidence limits for differences and ratios of two proportions and their odds ratios. *J. Amer. Statist. Assoc.* **72** 73–76.

WALTER, S. D. (1975). The distribution of Levin's measure of attributable risk. *Biometrika* **62** 371–372.

WALTER, S. D. (1976). The estimation and interpretation of attributable risk in health research. *Biometrics* **32** 829–849.

WANG, H. and RANALLI, M. G. (2007). Low-rank smoothing splines on complicated domains. *Biometrics* **63** 209–217. MR2345591 https://doi.org/10.1111/j.1541-0420.2006.00674.x

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1110–1126
https://doi.org/10.1214/25-AOAS2019

A PÓLYA TREE MODELLING FRAMEWORK FOR BATCH-MARK DATA

BY IOANNIS ROTOUS[1,a], ALEX DIANA[2,b] AND ELENI MATECHOU[3,c]

[1]*Department of Statistical Science, University College London,* [a]*i.rotous@ucl.ac.uk*
[2]*School of Mathematics, Statistics and Actuarial Science, University of Essex,* [b]*ad23269@essex.ac.uk*
[3]*Statistical Ecology @ Kent, School of Mathematics, Statistics and Actuarial Science, University of Kent,* [c]*e.matechou@kent.ac.uk*

Wildlife population surveys typically consist of multiple sampling occasions, where individuals are followed over time, enabling estimation of population size and, in open populations, of entry and exit patterns. Batch-mark (BM) surveys, where newly sampled individuals are given the same marking, often unique for each sampling occasion or each sampling period but not for each individual, provide the only viable monitoring tool for many species of amphibians, birds and fish. Modelling BM data for open populations has proven more challenging than modelling data where individuals are uniquely marked, and approaches proposed in the literature thus far rely on approximate inference or do not scale well with the number of individuals, and do not readily extend to the joint modelling of different observation processes often employed in practice. In this paper we propose a novel approach for modelling BM data, by defining a bivariate grid for modelling the latent entry and exit patterns, as well as population size. We employ the Bayesian nonparametric Pólya Tree (PT) prior for defining a model on the grid cells, which enables exact and highly efficient Bayesian inference on the number of individuals in each cell and hence of the population size and the entry/exit pattern. Our approach scales with the number of sampling occasions, instead of the number of individuals, and allows us to easily write the likelihood function for BM data under different observation processes. We demonstrate our new PT batch mark (PTBM) approach using extensive simulations and two case studies, comparing its performance with two recently proposed approaches.

REFERENCES

COWEN, L. L. E., BESBEAS, P., MORGAN, B. J. T. and SCHWARZ, C. J. (2014). A comparison of abundance estimates from extended batch-marking and Jolly–Seber-type experiments. *Ecol. Evol.* **4** 210–218. https://doi.org/10.1002/ece3.899

COWEN, L. L. E., BESBEAS, P., MORGAN, B. J. T. and SCHWARZ, C. J. (2017). Hidden Markov models for extended batch data. *Biometrics* **73** 1321–1331. MR3744545 https://doi.org/10.1111/biom.12701

DAVIDSON, J. R., SUDIRMAN, R., WAHID, I., BASKIN, R. N., HASAN, H., ARFAH, A. M., NUR, N., HIDAYAT, M. Y., SYAFRUDDIN, D. and LOBO, N. F. (2019). Mark-release-recapture studies reveal preferred spatial and temporal behaviors of Anopheles barbirostris in West Sulawesi, Indonesia. *Parasites Vectors* **12** 1–11.

DIANA, A., GRIFFIN, J. and MATECHOU, E. (2019). A Polya tree based model for unmarked individuals in an open wildlife population. In *Bayesian Statistics and New Generations*. *Springer Proc. Math. Stat.* **296** 3–11. Springer, Cham. MR4043141 https://doi.org/10.1007/978-3-030-30611-3_1

DIANA, A., MATECHOU, E., GRIFFIN, J., ARNOLD, T., TENAN, S. and VOLPONI, S. (2023). A general modeling framework for open wildlife populations based on the Polya tree prior. *Biometrics* **79** 2171–2183. MR4643985 https://doi.org/10.1111/biom.13756

DOLL, J. C., WOOD, C. J., GOODFRED, D. W. and RASH, J. M. (2021). Incorporating batch mark–recapture data into an integrated population model of Brown trout. *North Amer. J. Fish. Manag.* **41** 1390–1407.

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

Key words and phrases. Bayesian nonparametrics, stopover model, population size, removal models, survival probability.

HOLMES, C. C., CARON, F., GRIFFIN, J. E. and STEPHENS, D. A. (2015). Two-sample Bayesian nonparametric hypothesis testing. *Bayesian Anal*. **10** 297–320. MR3420884 https://doi.org/10.1214/14-BA914

HUGGINS, R., WANG, Y. and KEARNS, J. (2010). Analysis of an extended batch marking experiment using estimating equations. *J. Agric. Biol. Environ. Stat*. **15** 279–289. MR2787259 https://doi.org/10.1007/s13253-010-0018-6

JOLLY, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika* **52** 225–247. MR0210227 https://doi.org/10.1093/biomet/52.1-2.225

KENDALL, W. L. and POLLOCK, K. H. (1992). The robust design in capture-recapture studies: A review and evaluation by Monte Carlo simulation. *Wildlife* 2001: *Popul*. 31–43.

KING, R. and MCCREA, R. (2019). Capture-recapture methods and models: Estimating population size. In *Integrated Population Biology and Modeling*. *Part B*. *Handbook of Statist*. **40** 33–83. Elsevier/North-Holland, Amsterdam. MR3890221

LAVINE, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist*. **20** 1222–1235. MR1186248 https://doi.org/10.1214/aos/1176348767

MATECHOU, E. and CARON, F. (2017). Modelling individual migration patterns using a Bayesian nonparametric approach for capture-recapture data. *Ann. Appl. Stat*. **11** 21–40. MR3634313 https://doi.org/10.1214/16-AOAS989

MATECHOU, E., MCCREA, R. S., MORGAN, B. J. T., NASH, D. J. and GRIFFITHS, R. A. (2016). Open models for removal data. *Ann. Appl. Stat*. **10** 1572–1589. MR3553236 https://doi.org/10.1214/16-AOAS949

MCCREA, R. S. and MORGAN, B. J. T. (2015). *Analysis of Capture-Recapture Data*. *Chapman & Hall/CRC Interdisciplinary Statistics Series*. CRC Press, Boca Raton, FL. MR3330977

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

ROBERT, C. P. and CASELLA, G. (2004). The Metropolis—Hastings algorithm. In *Monte Carlo Statistical Methods* 267–320. Springer, New York, NY.

ROSSER, E., WILLDEN, S. A. and LOEB, G. M. (2022). Effects of SmartWater, a fluorescent mark, on the dispersal, behavior, and biocontrol efficacy of Phytoseiulus persimilis. *Exp. Appl. Acarol*. **87** 163–174. https://doi.org/10.1007/s10493-022-00732-9

ROTOUS, I., DIANA, A. and MATECHOU, E. (2025). Supplement to "A Pólya Tree modelling framework for batch-mark data." https://doi.org/10.1214/25-AOAS2019SUPPA, https://doi.org/10.1214/25-AOAS2019SUPPB

ROYLE, J. A., CONVERSE, S. J. and LINK, W. A. (2012). Data augmentation for hierarchical capture-recapture models. ArXiv preprint. Available at arXiv:1211.5706.

SCHWARZ, C. J. and ARNASON, A. N. (1996). A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics* **52** 860–873. MR1411736 https://doi.org/10.2307/2533048

SEBER, G. A. F. (1965). A note on the multiple-recapture census. *Biometrika* **52** 249–259. MR0210228 https://doi.org/10.1093/biomet/52.1-2.249

SEBER, G. A. F. and SCHOFIELD, M. R. (2019). *Capture-Recapture*: *Parameter Estimation for Open Animal Populations*. *Statistics for Biology and Health*. Springer, Cham. MR3967717 https://doi.org/10.1007/978-3-030-18187-1

VAVASSORI, L., SADDLER, A. and MÜLLER, P. (2019). Active dispersal of Aedes albopictus: A mark-release-recapture study using self-marking units. *Parasites Vectors* **12** 1–14.

ZHANG, W., BONNER, S. J. and MCCREA, R. S. (2023). Latent multinomial models for extended batch-mark data. *Biometrics* **79** 2732–2742. MR4644028 https://doi.org/10.1111/biom.13789

ZHANG, W., BRAVINGTON, M. V. and FEWSTER, R. M. (2019). Fast likelihood-based inference for latent count models using the saddlepoint approximation. *Biometrics* **75** 723–733. MR4012079 https://doi.org/10.1111/biom.13030

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1127–1146
https://doi.org/10.1214/25-AOAS2027

SYNTHESIZING DATA PRODUCTS, MATHEMATICAL MODELS, AND OBSERVATIONS FOR LAKE TEMPERATURE FORECASTING

BY MAIKE F. HOLTHUIJZEN[1,a], ROBERT B. GRAMACY[2,b], CAYELAN C. CAREY[3,d], DAVID M. HIGDON[2,c] AND R. QUINN THOMAS[4,e]

[1]*Virginia Tech & Sandia National Labs Dept 8738,* [a]*mholthuijzen@alumni.uidaho.edu*
[2]*Department of Statistics, Virginia Tech,* [b]*rbg@vt.edu,* [c]*dhigdon@vt.edu*
[3]*Department of Biological Sciences, Virginia Tech,* [d]*cayelan@vt.edu*
[4]*Departments of Forest Resources & Environmental Conservation and Biological Sciences, Virginia Tech,* [e]*rqthomas@vt.edu*

We present a novel forecasting framework for lake water temperature, which is crucial for managing lake ecosystems and drinking water resources. The General Lake Model (GLM) has been previously used for this purpose, but, similar to many process-based simulation models, it requires a large number of inputs (many of which are stochastic), presents challenges for uncertainty quantification (UQ), and can exhibit model bias. To address these issues, we propose a Gaussian process (GP) surrogate-based forecasting approach that efficiently handles large, high-dimensional data and accounts for input-dependent variability and systematic GLM bias. We validate the proposed approach and compare it with other forecasting methods, including a climatological model and raw GLM simulations. Our results demonstrate that our bias-corrected GP surrogate (GPBC) can outperform competing approaches in terms of forecast accuracy and UQ up to two weeks into the future.

REFERENCES

ANKENMAN, B., NELSON, B. L. and STAUM, J. (2008). Stochastic Kriging for simulation metamodeling. In 2008 *Winter Simulation Conference* 362–370. IEEE.

BAKER, E., BARBILLON, P., FADIKAR, A. et al. (2022). Analyzing stochastic computer models: A review with opportunities. *Statist. Sci.* **37** 64–89. MR4371096 https://doi.org/10.1214/21-sts822

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press/CRC, Boca Raton.

BINOIS, M. and GRAMACY, R. B. (2021). hetgp: Heteroskedastic Gaussian process modeling and sequential design in R. *J. Stat. Softw.* **98** 1–44.

BINOIS, M., GRAMACY, R. B. and LUDKOVSKI, M. (2018). Practical heteroscedastic Gaussian process modeling for large simulation experiments. *J. Comput. Graph. Statist.* **27** 808–821. MR3890872 https://doi.org/10.1080/10618600.2018.1458625

CAREY, C. C., IBELINGS, B. W., HOFFMANN, E. P., HAMILTON, D. P. and BROOKES, J. D. (2012). Ecophysiological adaptations that favour freshwater cyanobacteria in a changing climate. *Water Res.* **46** 1394–1407.

CAREY, C. C., WOELMER, W. M., LOFTON, M. E., FIGUEIREDO, R. J., BOOKOUT, B. J., CORRIGAN, R. S., DANESHMAND, V., HOUNSHELL, A. G., HOWARD, D. W. et al. (2022). Advancing lake and reservoir water quality management with near-term, iterative ecological forecasting. *Inland Waters* **12** 107–120.

FER, I., KELLY, R., MOORCROFT, P. R., RICHARDSON, A. D., COWDERY, E. M. and DIETZE, M. C. (2018). Linking big models to big data: Efficient ecosystem model calibration through Bayesian model emulation. *Biogeosciences* **15** 5801–5830.

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

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

Key words and phrases. Surrogate modeling, Gaussian process, Vecchia approximation, stochastic kriging.

GUINNESS, J. (2018). Permutation and grouping methods for sharpening Gaussian process approximations. *Technometrics* **60** 415–429. MR3878098 https://doi.org/10.1080/00401706.2018.1437476

HAMILL, T. M., WHITAKER, J. S., SHLYAEVA, A., BATES, G., FREDRICK, S., PEGION, P., SINSKY, E., ZHU, Y., TALLAPRAGADA, V. et al. (2022). The reanalysis for the global ensemble forecast system, version 12. *Mon. Weather Rev.* **150** 59–79.

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

HIPSEY, M. R., BRUCE, L. C., BOON, C., BUSCH, B., CAREY, C. C., HAMILTON, D. P., HANSON, P. C., READ, J. S., DE SOUSA, E. et al. (2019). A General Lake Model (GLM 3.0) for linking with high-frequency sensor data from the Global Lake Ecological Observatory Network (GLEON). *Geosci. Model Dev.* **12** 473–523.

HOLTHUIJZEN, M. F., GRAMACY, R. B., CAREY, C. C., HIGDON, D. M. and THOMAS, R. Q. (2025). Supplement to "Synthesizing data products, mathematical models, and observations for lake temperature forecasting." https://doi.org/10.1214/25-AOAS2027SUPPA, https://doi.org/10.1214/25-AOAS2027SUPPB

JOHNSON, L. R., GRAMACY, R. B., COHEN, J., MORDECAI, E., MURDOCK, C., ROHR, J., RYAN, S. J., STEWART-IBARRA, A. M. and WEIKEL, D. (2018). Phenomenological forecasting of disease incidence using heteroskedastic Gaussian processes: A Dengue case study. *Ann. Appl. Stat.* **12** 27–66. MR3773385 https://doi.org/10.1214/17-AOAS1090

KATZFUSS, M., GUINNESS, J. and LAWRENCE, E. (2022). Scaled Vecchia approximation for fast computer-model emulation. *SIAM/ASA J. Uncertain. Quantificat.* **10** 537–554. MR4490303 https://doi.org/10.1137/20M1352156

KATZFUSS, M., JUREK, M., ZILBER, D., GONG, W., GUINNESS, J., ZHANG, J. and SCHÄFER, F. (2020). GPvecchia: Fast Gaussian-process inference using Vecchia approximations. R package version 0.1.3.

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

LOFTON, M. E., HOWARD, D. W., THOMAS, R. Q. and CAREY, C. C. (2023). Progress and opportunities in advancing near-term forecasting of freshwater quality. *Glob. Change Biol.* **29** 1691–1714.

OAKLEY, J. E. and O'HAGAN, A. (2004). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 751–769. MR2088780 https://doi.org/10.1111/j.1467-9868.2004.05304.x

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

RAZAVI, S., TOLSON, B. A. and BURN, D. H. (2012). Review of surrogate modeling in water resources. *Water Resour. Res.* **48**.

ROY, P. T., EL MOÇAYD, N., RICCI, S., JOUHAUD, J.-C., GOUTAL, N., DE LOZZO, M. and ROCHOUX, M. C. (2018). Comparison of polynomial chaos and Gaussian process surrogates for uncertainty quantification and correlation estimation of spatially distributed open-channel steady flows. *Stoch. Environ. Res. Risk Assess.* **32** 1723–1741.

SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2018). *The Design and Analysis of Computer Experiments*. *Springer Series in Statistics*. Springer, New York. Second edition of [MR2160708]. MR3887662

THOMAS, R. Q., FIGUEIREDO, R. J., DANESHMAND, V., BOOKOUT, B. J., PUCKETT, L. K. and CAREY, C. C. (2020). A near-term iterative forecasting system successfully predicts reservoir hydrodynamics and partitions uncertainty in real time. *Water Resour. Res.* **56** e2019WR026138.

THOMAS, R. Q., MCCLURE, R. P., MOORE, T. N., WOELMER, W. M., BOETTIGER, C., FIGUEIREDO, R. J., HENSLEY, R. T. and CAREY, C. C. (2023). Near-term forecasts of NEON lakes reveal gradients of environmental predictability across the US. *Front. Ecol. Environ.* **21** 220–226.

WETZEL, R. G. (2001). *Limnology*: *Lake and River Ecosystems*. Gulf Pub., Houston.

YANG, J., JAKEMAN, A., FANG, G. and CHEN, X. (2018). Uncertainty analysis of a semi-distributed hydrologic model based on a Gaussian process emulator. *Environ. Model. Softw.* **101** 289–300.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1147–1166
https://doi.org/10.1214/25-AOAS2035

A DUAL-DICTIONARY MODEL FOR MINING DOMAIN-SPECIFIC CHINESE TEXTS

BY JIAZE XU[1,a], CHANGZAI PAN[1,b] AND KE DENG[2,c]

[1]*Center for Statistical Science and Department of Industrial Engineering, Tsinghua University,* [a]*xjz16@tsinghua.org.cn,* [b]*pcz18@tsinghua.org.cn*

[2]*Department of Statistics and Data Science, Tsinghua University,* [c]*kdeng@tsinghua.edu.cn*

Processing domain-specific Chinese texts is one of the most challenging tasks in *natural language processing* (NLP) due to the unique characteristics of Chinese and specific challenges in processing domain-specific texts. Existing NLP methods, based on supervised learning or large language models, often suffer from unstable performance in processing domain-specific Chinese texts. In this study we propose a novel statistical approach called TopWORDS-MEPA (Abbreviated as TWM) that can achieve high-quality meta-pattern discovery, named entity recognition, text segmentation, and relation extraction simultaneously from unannotated target texts with little training information. Simulation studies and real data applications confirm that TopWORDS-MEPA is a powerful alternative of existing NLP methods for processing domain-specific Chinese texts that enjoys competitive performance, transparent interpretation, low training and computing costs, and efficient utilization of domain knowledge for processing domain-specific Chinese texts.

REFERENCES

BAHL, L. R., JELINEK, F. and MERCER, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **5** 179–190. https://doi.org/10.1109/tpami.1983.4767370

BAUM, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In *Inequalities, III* (*Proc. Third Sympos., Univ. California, Los Angeles, Calif.*, 1969; *Dedicated to the Memory of Theodore S. Motzkin*) 1–8. Academic Press, New York. MR0341782

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

BROWN, T., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J. D., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G. et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, eds.) **33** 1877–1901. Curran Associates, Red Hook.

CHE, W., FENG, Y., QIN, L. and LIU, T. (2020). N-LTP: A open-source neural Chinese language technology platform with pretrained models. Available at arXiv:2009.11616 [cs].

CHEN, S. F. and GOODMAN, J. (1999). An empirical study of smoothing techniques for language modeling. *Comput. Speech Lang.* **13** 359–394.

CUI, Y., CHE, W., LIU, T., QIN, B., YANG, Z., WANG, S. and HU, G. (2019). Pre-training with whole word masking for Chinese BERT. *CoRR* Available at arXiv:1906.08101.

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

DENG, K., BOL, P. K., LI, K. J. and LIU, J. S. (2016). On the unsupervised analysis of domain-specific Chinese texts. *Proc. Natl. Acad. Sci. USA* **113** 6154–6159. MR3514612 https://doi.org/10.1073/pnas.1516510113

DU, Z., QIAN, Y., LIU, X., DING, M., QIU, J., YANG, Z. and TANG, J. (2022). GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the* 60*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*) 320–335.

DUAN, H. and ZHENG, Y. (2011). A study on features of the CRFs-based Chinese named entity recognition. *Int. J. Adv. Intel. Paradig.* **3** 287–294.

Key words and phrases. Natural language processing, meta-pattern recognition, named entity recognition, rare word discovery, Bayesian inference.

GENG, Z., YAN, H., QIU, X. and HUANG, X. (2020). fastHan: A BERT-based joint many-task toolkit for Chinese NLP. Available at arXiv:2009.08633 [cs].

JIANG, A. Q., SABLAYROLLES, A., MENSCH, A., BAMFORD, C., CHAPLOT, D. S., CASAS, D., BRESSAND, F., LENGYEL, G., LAMPLE, G. et al. (2023). Mistral 7B. arXiv preprint. Available at arXiv:2310.06825.

JIANG, M., SHANG, J., CASSIDY, T., REN, X., KAPLAN, L. M., HANRATTY, T. P. and HAN, J. (2017). MetaPAD: Meta pattern discovery from massive text corpora. In *Proceedings of the* 23*rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 877–886.

LAFFERTY, J. D., MCCALLUM, A. and PEREIRA, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. *ICML* '01 282–289. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

LUO, R., XU, J., ZHANG, Y., REN, X. and SUN, X. (2019). PKUSEG: A toolkit for multi-domain Chinese word segmentation. Available at arXiv:1906.11455 [cs].

PAN, C., LI, F. and DENG, K. (2023). TopWORDS-Poetry: Simultaneous text segmentation and word discovery for classical Chinese poetry via Bayesian inference. In *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing* (H. Bouamor, J. Pino and K. Bali, eds.) 3372–3386. Association for Computational Linguistics, Singapore.

PAN, C., SUN, M. and DENG, K. (2022). TopWORDS-Seg: Simultaneous text segmentation and word discovery for open-domain Chinese texts via Bayesian inference. In *Proceedings of the* 60*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*) (S. Muresan, P. Nakov and A. Villavicencio, eds.) 158–169. Association for Computational Linguistics, Dublin, Ireland.

PENG, F., FENG, F. and MCCALLUM, A. (2004). Chinese segmentation and new word detection using conditional random fields. In *COLING* 2004: *Proceedings of the* 20*th International Conference on Computational Linguistics* 562–568. COLING, Geneva, Switzerland.

QI, P., ZHANG, Y., ZHANG, Y., BOLTON, J. and MANNING, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*: *System Demonstrations* 101–108. Assoc. Comput. Linguistics.

SHIEH, Y.-P. (2011). Appositional term clip: A subject-oriented appositional term extraction algorithm. In *New Eyes for Discovery*: *Foundations and Imaginations of Digital Humanities* 133–162.

SUN, J. (2012). Jieba Chinese text segmentation.

SUN, M., CHEN, X., ZHANG, K., GUO, Z. and LIU, Z. (2016). THULAC: An efficient lexical analyzer for Chinese.

SUN, Z. and DENG, Z.-H. (2018). Unsupervised neural word segmentation for Chinese via segmental language modeling. In *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing* 4915–4920. Association for Computational Linguistics, Brussels, Belgium.

TOUVRON, H., LAVRIL, T., IZACARD, G., MARTINET, X., LACHAUX, M.-A., LACROIX, T., ROZIÈRE, B., GOYAL, N., HAMBRO, E. et al. (2023). Llama: Open and efficient foundation language models (2023). arXiv preprint. Available at arXiv:2302.13971.

VITERBI, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans*. *Inf. Theory* **13** 260–269.

WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann*. *Math*. *Stat*. **9** 60–62.

XU, J., PAN, C. and DENG, K. (2025). Supplement to "A dual-dictionary model for mining domain-specific Chinese texts." https://doi.org/10.1214/25-AOAS2035SUPPA, https://doi.org/10.1214/25-AOAS2035SUPPB

ZHANG, Y., XU, Z. and ZHANG, T. (2008). Fusion of multiple features for Chinese named entity recognition based on CRF model. In *Proceedings of the* 4*th Asia Information Retrieval Conference on Information Retrieval Technology*. *AIRS*'08 95–106. Springer, Berlin, Heidelberg.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1167–1189
https://doi.org/10.1214/24-AOAS1955

KULLBACK-LEIBLER-BASED DISCRETE FAILURE TIME MODELS FOR INTEGRATION OF PUBLISHED PREDICTION MODELS WITH NEW TIME-TO-EVENT DATASET

BY DI WANG[1,a], WEN YE[1,b], RANDALL SUNG[2,f], HUI JIANG[1,c], JEREMY M. G. TAYLOR[1,d], LISA LY[3,g] AND KEVIN HE[1,e]

[1]*Department of Biostatistics, University of Michigan,* [a]*dwwang@umich.edu,* [b]*wye@umich.edu,* [c]*jianghui@umich.edu,* [d]*jmgt@umich.edu,* [e]*kevinhe@umich.edu*

[2]*Section of Transplantation Surgery, University of Michigan,* [f]*rssung@med.umich.edu*

[3]*Department of Urology, Temple University,* [g]*lylisa2@gmail.com*

Prediction of time-to-event data often suffers from rare event rates, small sample sizes, high dimensionality, and low signal-to-noise ratios. Incorporating published prediction models from external large-scale studies is expected to improve the performance of prognosis prediction from internal individual-level data. However, existing integration approaches typically assume that the underlying distributions of the external and internal data sources are similar, which is often invalid. To account for challenges, including heterogeneity, data sharing, and privacy constraints, we propose a failure time integration procedure, which utilizes a discrete hazard-based Kullback–Leibler discriminatory information measuring the discrepancy between the external models and the internal dataset. The asymptotic properties and simulation results show the advantage of the proposed method compared to those solely based on internal data. We apply the proposed method to improve prediction performance on a kidney transplant dataset from a local hospital by integrating this small-sized dataset with a published survival model obtained from the national transplant registry.

REFERENCES

AREF, A., SHARMA, A. and HALAWA, A. (2017). Smoking in renal transplantation; facts beyond myth. *World J. Transplant*. **7** 129–133. https://doi.org/10.5500/wjt.v7.i2.129

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

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

BÜHLMANN, P. and YU, B. (2003). Boosting with the L_2 loss: Regression and classification. *J. Amer. Statist. Assoc*. **98** 324–339. MR1995709 https://doi.org/10.1198/016214503000125

BURNHAM, K. P. and ANDERSON, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer, New York. MR1919620

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

CHAUDHURI, S., HANDCOCK, M. S. and RENDALL, M. S. (2008). Generalized linear models incorporating population level information: An empirical-likelihood-based approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **70** 311–328. MR2424755 https://doi.org/10.1111/j.1467-9868.2007.00637.x

CHEN, J. H. and QIN, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika* **80** 107–116. MR1225218 https://doi.org/10.1093/biomet/80.1.107

CHEN, Z., NING, J., SHEN, Y. and QIN, J. (2021). Combining primary cohort data with external aggregate information without assuming comparability. *Biometrics* **77** 1024–1036. MR4320675 https://doi.org/10.1111/biom.13356

Key words and phrases. Data integration, kidney transplant, relative entropy, survival prediction.

CHENG, Y.-J., LIU, Y.-C., TSAI, C.-Y. and HUANG, C.-Y. (2023). Semiparametric estimation of the transformation model by leveraging external aggregate data in the presence of population heterogeneity. *Biometrics* **79** 1996–2009. MR4643972 https://doi.org/10.1111/biom.13778

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

DEBRAY, T. P. A., KOFFIJBERG, H., NIEBOER, D., VERGOUWE, Y., STEYERBERG, E. W. and MOONS, K. G. M. (2014). Meta-analysis and aggregation of multiple published prediction models. *Stat. Med.* **33** 2341–2362. MR3256671 https://doi.org/10.1002/sim.6080

DING, J., LI, J., HAN, Y., MCKEAGUE, I. W. and WANG, X. (2023). Fitting additive risk models using auxiliary information. *Stat. Med.* **42** 894–916. MR4550102 https://doi.org/10.1002/sim.9649

ESTES, J. P., MUKHERJEE, B. and TAYLOR, J. M. G. (2017). Empirical Bayes estimation and prediction using summary-level information from external big data sources adjusting for violations of transportability. *Stat. Biosci.* **10** 568–586.

GENSHEIMER, M. F. and NARASIMHAN, B. (2019). A scalable discrete-time survival model for neural networks. *PeerJ* **7** e6257. https://doi.org/10.7717/peerj.6257

HAN, B., VAN KEILEGOM, I. and WANG, X. (2022). Semiparametric estimation of the nonmixture cure model with auxiliary survival information. *Biometrics* **78** 448–459. MR4450567 https://doi.org/10.1111/biom.13450

HAN, P. and LAWLESS, J. F. (2019). Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statist. Sinica* **29** 1321–1342. MR3932520

HARIHARAN, S., ISRANI, A. K. and DANOVITCH, G. (2021). Long-term survival after kidney transplantation. *N. Engl. J. Med.* **385** 729–743. https://doi.org/10.1056/NEJMra2014530

HE, K., ZHU, J., KANG, J. and LI, Y. (2022). Stratified Cox models with time-varying effects for national kidney transplant patients: A new blockwise steepest ascent method. *Biometrics* **78** 1221–1232. MR4493519 https://doi.org/10.1111/biom.13473

HECTOR, E. C. and MARTIN, R. (2024). Turning the information-sharing dial: Efficient inference from different data sources. *Electron. J. Stat.* **18** 2974–3020. MR4772516 https://doi.org/10.1214/24-ejs2265

HJORT, N. L. and POLLARD, D. (1993). Asymptotics for minimisers of convex processes. Available at arXiv:1107.3806.

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, Y., HE, Y. and ZHANG, H. (2016). Variable selection with prior information for generalized linear models via the prior LASSO method. *J. Amer. Statist. Assoc.* **111** 355–376. MR3494665 https://doi.org/10.1080/01621459.2015.1008363

JOHANSEN, K. L., CHERTOW, G. M., GILBERTSON, D. T., ISHANI, A., ISRANI, A. K., KU, E., LI, S. et al. (2023). US renal data system 2022 annual data report: Epidemiology of kidney disease in the United States. *Amer. J. Kidney Dis.* **81** A8–A11.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* **60** 267–278. MR0326939 https://doi.org/10.1093/biomet/60.2.267

KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR1924807 https://doi.org/10.1002/9781118032985

KASISKE, B. L., WEY, A., SALKOWSKI, N., ZAUN, D., SCHAFFHAUSEN, C. R., ISRANI, A. K. and SNYDER, J. J. (2019). Seeking new answers to old questions about public reporting of transplant program performance in the United States. *Amer. J. Transplant.* **19** 317–323.

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

LIOUDAKI, E., ANDROVITSANEA, A., PETRAKIS, I., BAKOGIANNIS, C. and ANDROULAKIS, E. (2022). Cardiac imaging and management of cardiac disease in asymptomatic renal transplant candidates: A current update. *Diagnostics* **12**. https://doi.org/10.3390/diagnostics12102332

LIU, C. and SHUM, H. Y. (2003). Kullback–Leibler boosting. In 2003 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003. *Proceedings* **1** I–I.

LIU, D., ZHENG, Y., PRENTICE, R. L. and HSU, L. (2014). Estimating risk with time-to-event data: An application to the women's health initiative. *J. Amer. Statist. Assoc.* **109** 514–524. MR3223729 https://doi.org/10.1080/01621459.2014.881739

MAYR, A., HOFNER, B. and SCHMID, M. (2012). The importance of knowing when to stop. A sequential stopping rule for component-wise gradient boosting. *Methods Inf. Med.* **51** 178–186. https://doi.org/10.3414/ME11-02-0030

PARAMESH, A. S., DAVIS, J. Y., MALLIKARJUN, C., ZHANG, R., CANNON, R., SHORES, N., KILLACKEY, M. T., MCGEE, J., SAGGI, B. H. et al. (2012). Kidney transplantation alone in ESRD patients with hepatitis C cirrhosis. *Transplantation* **94** 250–254. https://doi.org/10.1097/TP.0b013e318255f890

PARKER, R., ARMSTRONG, M. J., CORBETT, C., DAY, E. J. and NEUBERGER, J. M. (2013). Alcohol and substance abuse in solid-organ transplant recipients. *Transplantation* **96** 1015–1024. https://doi.org/10.1097/TP.0b013e31829f7579

POGGIO, E. D., AUGUSTINE, J. J., ARRIGAIN, S., BRENNAN, D. C. and SCHOLD, J. D. (2021). Long-term kidney transplant graft survival-making progress when most needed. *Amer. J. Transplant*. **21** 2824–2832. https://doi.org/10.1111/ajt.16463

PRENTICE, R. L. and KALBFLEISCH, J. D. (2003). Mixed discrete and continuous Cox regression model. *Lifetime Data Anal*. **9** 195–210. MR1970405 https://doi.org/10.1023/A:1022935019768

QIN, J. (2000). Combining parametric and empirical likelihoods. *Biometrika* **87** 484–490. MR1782493 https://doi.org/10.1093/biomet/87.2.484

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

SCHAPIRE, R., ROCHERY, M., GILBERT, M. and GUPTA, N. (2005). Boosting with prior knowledge for call classification. *IEEE Trans*. *Speech Audio Process*. **13** 174–181.

SHENG, Y., SUN, Y., HUANG, C.-Y. and KIM, M.-O. (2021). Synthesizing external aggregated information in the penalized Cox regression under population heterogeneity. *Stat*. *Med*. **40** 4915–4930. MR4315459 https://doi.org/10.1002/sim.9101

SHENG, Y., SUN, Y., HUANG, C.-Y. and KIM, M.-O. (2022). Synthesizing external aggregated information in the presence of population heterogeneity: A penalized empirical likelihood approach. *Biometrics* **78** 679–690. MR4450586 https://doi.org/10.1111/biom.13429

SNYDER, J. J., SALKOWSKI, N., KIM, S. J., ZAUN, D., XIONG, H., ISRANI, A. K. and KASISKE, B. L. (2016). Developing statistical models to assess transplant outcomes using national registries: The process in the United States. *Transplantation* **100** 288–294. https://doi.org/10.1097/TP.0000000000000891

STALLONE, G., PONTRELLI, P., RASCIO, F., CASTELLANO, G., GESUALDO, L. and GRANDALIANO, G. (2020). Coagulation and fibrinolysis in kidney graft rejection. *Front*. *Immunol*. **11** 1807. https://doi.org/10.3389/fimmu.2020.01807

TUTZ, G. and SCHMID, M. (2016). *Modeling Discrete Time-to-Event Data*. *Springer Series in Statistics*. Springer, Cham. MR3497009 https://doi.org/10.1007/978-3-319-28158-2

WANG, D., YE, W., SUNG, R., JIANG, H., TAYLOR, J. M., LY, L. and HE, K. (2025). Supplement to "Kullback-Leibler-based Discrete Failure Time Models for Integration of Published Prediction Models with New Time-to-Event Dataset." https://doi.org/10.1214/24-AOAS1955SUPPA, https://doi.org/10.1214/24-AOAS1955SUPPB

WOLFE, R. A., ASHBY, V. B., MILFORD, E. L., OJO, A. O., ETTENGER, R. E., AGODOA, L. Y., HELD, P. J. and PORT, F. K. (1999). Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *N*. *Engl*. *J*. *Med*. **341** 1725–1730.

ZHAI, Y. and HAN, P. (2022). Data integration with oracle use of external information from heterogeneous populations. *J*. *Comput*. *Graph*. *Statist*. **31** 1001–1012. MR4513365 https://doi.org/10.1080/10618600.2022.2050248

ZHANG, T. and YU, B. (2005). Boosting with early stopping: Convergence and consistency. *Ann*. *Statist*. **33** 1538–1579. MR2166555 https://doi.org/10.1214/009053605000000255

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1190–1213
https://doi.org/10.1214/24-AOAS1982

HOLDOUT SETS FOR SAFE PREDICTIVE MODEL UPDATING

BY SAMI HAIDAR-WEHBE[a], SAMUEL R. EMERSON[b], LOUIS J. M. ASLETT[c] AND JAMES LILEY[d]

Department of Mathematical Sciences, Durham University, [a]*sami.haidar96@gmail.com,* [b]*samuel.emerson@hotmail.co.uk,* [c]*louis.aslett@durham.ac.uk,* [d]*james.liley@durham.ac.uk*

Predictive risk scores for adverse outcomes are increasingly crucial in guiding health interventions. Such scores may need to be periodically updated due to change in the distributions they model. However, directly updating risk scores used to guide intervention can lead to biased risk estimates. To address this, we propose updating using a "holdout set", a subset of the population that does not receive interventions guided by the risk score. Balancing the holdout set size is essential to ensure good performance of the updated risk score while minimising the number of held out samples. We prove that this approach reduces adverse outcome frequency to an asymptotically optimal level and argue that often there is no competitive alternative. We describe conditions under which an optimal holdout size (OHS) can be readily identified and introduce parametric and semiparametric algorithms for OHS estimation. We apply our methods to the ASPRE risk score for pre-eclampsia to recommend a plan for updating it in the presence of change in the underlying data distribution. We show that, in order to minimise the number of pre-eclampsia cases over time, this is best achieved using a holdout set of around 10,000 individuals.

REFERENCES

ACOG (2016). Practice advisory on low-dose aspirin and prevention of preeclampsia: Updated recommendations. American College of Obstetricians and Gynecologists (ACOG).

AKOLEKAR, R., SYNGELAKI, A., POON, L., WRIGHT, D. and NICOLAIDES, K. H. (2013). Competing risks model in early screening for preeclampsia by biophysical and biochemical markers. *Fetal Diagn*. *Ther.* **33** 8–15. https://doi.org/10.1159/000341264

ALAA, A. M. and VAN DER SCHAAR, M. (2018). Autoprognosis: Automated clinical prognostic modeling via Bayesian optimization with structured kernel learning. arXiv preprint arXiv:1802.07207.

AMARI, S.-I. (1993). A universal theorem on learning curves. *Neural Netw.* **6** 161–166.

BROCHU, E., CORA, V. M. and DE FREITAS, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint arXiv:1012.2599.

CHALMERS, J., PULLAN, M., FABRI, B., MCSHANE, J., SHAW, M., MEDIRATTA, N. and POULLIS, M. (2013). Validation of EuroSCORE II in a modern cohort of patients undergoing cardiac surgery. *Eur. J. Cardio-Thorac*. *Surg*. **43** 688–694.

CHISLETT, L., ASLETT, L. J. M., DAVIES, A. R., VALLEJOS, C. A. and LILEY, J. (2024). Ethical considerations of use of hold-out sets in clinical prediction model management. *AI Ethics*. https://doi.org/10.1007/s43681-024-00561-z

COLLINS, G. S., REITSMA, J. B., ALTMAN, D. G. and MOONS, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Circulation* **131** 211–219.

COOK, J. A. and COLLINS, G. S. (2015). The rise of big clinical databases. *Br. J. Surg*. **102** e93–e101. https://doi.org/10.1002/bjs.9723

DE LANGE, M. and TUYTELAARS, T. (2021). Continual prototype evolution: Learning online from non-stationary data streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 8250–8259.

DRUSVYATSKIY, D. and XIAO, L. (2023). Stochastic optimization with decision-dependent distributions. *Math*. *Oper. Res*. **48** 954–998. MR4588947

Key words and phrases. Model updating, performative prediction, updating paradox.

FINLAYSON, S. G., SUBBASWAMY, A., SINGH, K., BOWERS, J., KUPKE, A., ZITTRAIN, J., KOHANE, I. S. and SARIA, S. (2020). The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* 283–286.

HAIDAR-WEHBE, S., EMERSON, S., ASLETT, L. J. and LILEY, J. (2025a). Supplements S1–S10 to "Holdout sets for safe predictive model updating." https://doi.org/10.1214/24-AOAS1982SUPPA

HAIDAR-WEHBE, S., EMERSON, S., ASLETT, L. J. and LILEY, J. (2025b). Supplements S11 to "Holdout sets for safe predictive model updating." https://doi.org/10.1214/24-AOAS1982SUPPB

HIPPISLEY-COX, J., COUPLAND, C. and BRINDLE, P. (2017). Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: Prospective cohort study. *BMJ* **357** j2099. https://doi.org/10.1136/bmj.j2099

IZZO, Z., ZOU, J. and YING, L. (2021). How to learn when data gradually reacts to your model. arXiv preprint arXiv:2112.07042.

KANSAGARA, D., ENGLANDER, H., SALANITRO, A., KAGEN, D., THEOBALD, C., FREEMAN, M. and KRIPALANI, S. (2011). Risk prediction models for hospital readmission: A systematic review. *J. Amer. Med. Assoc.* **306** 1688–1698.

KRANZER, K., SIMMS, V., DAUYA, E., OLARU, I. D., DZIVA CHIKWARI, C., MARTIN, K., REDZO, N., BANDASON, T., TEMBO, M. et al. (2021). Identifying youth at high risk for sexually transmitted infections in community-based settings using a risk prediction tool: A validation study. *BMC Infect. Dis.* **21** 1–9.

LENERT, M. C., MATHENY, M. E. and WALSH, C. G. (2019). Prognostic models will be victims of their own success, unless.... *J. Amer. Med. Inform. Assoc.* **26** 1645–1650.

LI, Q. and WAI, H.-T. (2021). State dependent performative prediction with stochastic approximation. arXiv preprint arXiv:2110.00800.

LILEY, J. (2021). Stacking interventions for equitable outcomes. arXiv preprint arXiv:2110.04163.

LILEY, J., BOHNER, G., EMERSON, S. R., MATEEN, B. A., BORLAND, K., CARR, D., HEALD, S., ODURO, S. D., IRELAND, J. et al. (2024). Development and assessment of a machine learning tool for predicting emergency admission in Scotland. *npj Digit. Med.* **7** 277.

LILEY, J., EMERSON, S. R., MATEEN, B. A., VALLEJOS, C. A., ASLETT, L. J. M. and VOLLMER, S. J. (2021). Model updating after interventions paradoxically introduces bias. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*. *Proceedings of Machine Learning Research* **130** 3916–3924.

LU, J., LIU, A., DONG, F., GU, F., GAMA, J. and ZHANG, G. (2018). Learning under concept drift: A review. *IEEE Trans. Knowl. Data Eng.* **31** 2346–2363.

MENDLER-DÜNNER, C., PERDOMO, J. C., ZRNIC, T. and HARDT, M. (2020). Stochastic optimization for performative prediction. arXiv preprint arXiv:2006.06887.

NASHEF, S. A., ROQUES, F., SHARPLES, L. D., NILSSON, J., SMITH, C., GOLDSTONE, A. R. and LOCKOWANDT, U. (2012). EuroSCORE II. *Eur. J. Cardio-Thorac. Surg.* **41** 734–745.

O'GORMAN, N., WRIGHT, D., POON, L., ROLNIK, D. L., SYNGELAKI, A., DE ALVARADO, M., CARBONE, I. F., DUTEMEYER, V., FIOLNA, M. et al. (2017). Multicenter screening for pre-eclampsia by maternal factors and biomarkers at 11–13 weeks' gestation: Comparison with NICE guidelines and ACOG recommendations. *Ultrasound Obstet. Gynecol.* **49** 756–760.

OCHS, A., MCGURNAGHAN, S., BLACK, M. W., LEESE, G. P., PHILIP, S., SATTAR, N., STYLES, C., WILD, S. H., MCKEIGUE, P. M., COLHOUN, H. M. et al. (2019). Use of personalised risk-based screening schedules to optimise workload and sojourn time in screening programmes for diabetic retinopathy: A retrospective cohort study. *PLoS Med.* **16** e1002945.

PERDOMO, J., ZRNIC, T., MENDLER-DÜNNER, C. and HARDT, M. (2020). Performative prediction. In *International Conference on Machine Learning* 7599–7609. PMLR.

RILEY, R. D., ENSOR, J., SNELL, K. I. E., HARRELL JR., F. E., MARTIN, G. P., REITSMA, J. B., MOONS, K. G. M., COLLINS, G. and VAN SMEDEN, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ* **368** m441. https://doi.org/10.1136/bmj.m441

ROLNIK, D. L., WRIGHT, D., POON, L., SYNGELAKI, A., O'GORMAN, N., DE PACO MATALLANA, C., AKOLEKAR, R., CICERO, S., JANGA, D. et al. (2017a). ASPRE trial: Performance of screening for preterm pre-eclampsia. *Ultrasound Obstet. Gynecol.* **50** 492–495.

ROLNIK, D. L., WRIGHT, D., POON, L. C., O'GORMAN, N., SYNGELAKI, A., DE PACO MATALLANA, C., AKOLEKAR, R., CICERO, S., JANGA, D. et al. (2017b). Aspirin versus placebo in pregnancies at high risk for preterm preeclampsia. *N. Engl. J. Med.* **377** 613–622.

SHAHIAN, D. M., JACOBS, J. P., BADHWAR, V., KURLANSKY, P. A., FURNARY, A. P., CLEVELAND JR., J. C., LOBDELL, K. W., VASSILEVA, C., VON BALLMOOS, M. C. W. et al. (2018). The society of thoracic surgeons 2018 adult cardiac surgery risk models: Part 1—background, design considerations, and model development. *Ann. Thorac. Surg.* **105** 1411–1418.

SPERRIN, M., JENKINS, D., MARTIN, G. P. and PEEK, N. (2019). Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *J. Amer. Med. Inform. Assoc.* **26** 1675–1676. https://doi.org/10.1093/jamia/ocz197

STALLARD, N., MILLER, F., DAY, S., HEE, S. W., MADAN, J., ZOHAR, S. and POSCH, M. (2017). Determination of the optimal sample size for clinical trial accounting for the population size. *Biom. J.* **59** 609–625. MR3672685 https://doi.org/10.1002/bimj.201500228

TOPOL, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **25** 44–56. https://doi.org/10.1038/s41591-018-0300-7

TSYMBAL, A. (2004). The problem of concept drift: Definitions and related work. Computer Science Dept., Trinity College Dublin 106 58.

USFDA et al. (2019). Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD).

VIERING, T. and LOOG, M. (2021). The shape of learning curves: A review. arXiv preprint arXiv:2103.10948.

WALLACE, E., STUART, E., VAUGHAN, N., BENNETT, K., FAHEY, T. and SMITH, S. M. (2014). Risk prediction models to predict emergency hospital admission in community-dwelling adults: A systematic review. *Med. Care* **52** 751.

ZUKANOVIĆ, A. (2013). Caries risk assessment models in caries prediction. *Acta Med. Acad.* **42** 198–208. https://doi.org/10.5644/ama2006-124.87

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1214–1239
https://doi.org/10.1214/24-AOAS2006

MODELING STRUCTURE AND COUNTRY-SPECIFIC HETEROGENEITY IN MISCLASSIFICATION MATRICES OF VERBAL AUTOPSY-BASED CAUSE OF DEATH CLASSIFIERS

BY SANDIPAN PRAMANIK[1,a], SCOTT ZEGER[1,b], DIANNA BLAU[2,d] AND ABHIRUP DATTA[1,c]

[1]*Department of Biostatistics, Johns Hopkins University,* [a]*spraman4@jhmi.edu,* [b]*sz@jhu.edu,* [c]*abhidatta@jhu.edu*
[2]*Global Health Center, U.S. Centers for Disease Control and Prevention,* [d]*bvv1@cdc.gov*

Verbal autopsy (VA) algorithms are routinely used to determine individual-level causes of death (COD) in many low- and-middle-income countries. The individual CODs are then aggregated to derive population-level cause-specific mortality fractions (CSMF), which are essential to informing public health policies. However, VA algorithms frequently misclassify COD and introduce bias in CSMF estimates. A recent method, *VA-calibration*, can correct for this bias using a VA misclassification rate matrix estimated from paired data on COD from both VA and minimally invasive tissue sampling (MITS) from the Child Health and Mortality Prevention Surveillance (CHAMPS) Network. Due to the limited sample size, CHAMPS data are pooled across all countries, implicitly assuming that the misclassification rates are homogeneous.

In this research we show that the VA misclassification matrices are substantially heterogeneous across countries, thereby biasing the VA-calibration. We develop a coherent framework for modeling country-specific VA misclassification matrices in data-scarce settings. We first introduce a novel *base model* to parsimoniously characterize misclassifications via two latent mechanisms—*intrinsic accuracy* and *systematic preference*. We prove that these mechanisms are identifiable from the data and manifest as a form of invariance in certain misclassification odds, a pattern evident in the CHAMPS data. Then we expand from this base model, adding higher complexity and country-specific heterogeneity via interpretable effect sizes. Shrinkage priors balance the bias-variance trade-off by adaptively favoring simpler models. We publish uncertainty-quantified estimates of VA misclassification rates for six countries. This effort broadens VA-calibration's future applicability and strengthens ongoing efforts of using VA for mortality surveillance.

REFERENCES

ADAIR, T. and LOPEZ, A. D. (2018). Estimating the completeness of death registration: An empirical method. *PLoS ONE* **13** 1–19. https://doi.org/10.1371/journal.pone.0197047

BASSAT, Q., CASTILLO, P., MARTÍNEZ, M. J., JORDAO, D., LOVANE, L., HURTADO, J. C., NHAMPOSSA, T., SANTOS RITCHIE, P., BANDEIRA, S. et al. (2017). Validity of a minimally invasive autopsy tool for cause of death determination in pediatric deaths in Mozambique: An observational study. *PLoS Med*. **14** e1002317.

BEESLEY, L. J. and MUKHERJEE, B. (2022). Statistical inference for association studies using electronic health records: Handling both selection bias and outcome misclassification. *Biometrics* **78** 214–226. MR4408582 https://doi.org/10.1111/biom.13400

BLAU, D. M., CANEER, J. P., PHILIPSBORN, R. P., MADHI, S. A., BASSAT, Q., VARO, R., MANDOMANDO, I., IGUNZA, K. A., KOTLOFF, K. L. et al. (2019). Overview and development of the child health and mortality prevention surveillance determination of cause of death (DeCoDe) process and DeCoDe diagnosis standards. *Clin*. *Infect*. *Dis*. **69** S333–S341.

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

Key words and phrases. Bayesian, hierarchical modeling, verbal autopsy, minimally invasive tissue sampling, global health.

BYASS, P., CHANDRAMOHAN, D., CLARK, S. J., D'AMBRUOSO, L., FOTTRELL, E., GRAHAM, W. J., HERBST, A. J., HODGSON, A., HOUNTON, S. et al. (2012). Strengthening standardised interpretation of verbal autopsy data: The new InterVA-4 tool. *Glob. Health Action* **5** 19281. PMID: 28140847. https://doi.org/10.3402/gha.v5i0.19281

CARSHON-MARSH, R., AIMONE, A., ANSUMANA, R., SWARAY, I. B., ASSALIF, A., MUSA, A., MEH, C., SMART, F., FU, S. H. et al. (2022). Child, maternal, and adult mortality in Sierra Leone: Nationally representative mortality survey 2018–20. *Lancet Glob. Health* **10** e114–e123.

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

CHRISTIANSEN, C. L. and MORRIS, C. N. (1997). Hierarchical Poisson regression modeling. *J. Amer. Statist. Assoc.* **92** 618–632. MR1467853 https://doi.org/10.2307/2965709

CLARK, S. J., LI, Z. and MCCORMICK, T. H. (2018). Quantifying the contributions of training data and algorithm logic to the performance of automated cause-assignment algorithms for verbal autopsy. arXiv preprint arXiv:1803.07141.

CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Mach. Learn.* **20** 273–297.

DANIELS, M. J. (1999). A prior for the variance in hierarchical models. *Canad. J. Statist.* **27** 567–578. MR1745822 https://doi.org/10.2307/3316112

DATTA, A., FIKSEL, J., AMOUZOU, A. and ZEGER, S. L. (2021). Regularized Bayesian transfer learning for population-level etiological distributions. *Biostatistics* **22** 836–857. MR4325730 https://doi.org/10.1093/biostatistics/kxaa001

DE SAVIGNY, D., RILEY, I., CHANDRAMOHAN, D., ODHIAMBO, F., NICHOLS, E., NOTZON, S., ABOUZAHR, C., MITRA, R., COBOS MUÑOZ, D. et al. (2017). Integrating community-based verbal autopsy into civil registration and vital statistics (CRVS): System-level considerations. *Glob. Health Action* **10** 1272882.

FIKSEL, J., DATTA, A., AMOUZOU, A. and ZEGER, S. (2022). Generalized Bayes quantification learning under dataset shift. *J. Amer. Statist. Assoc.* **117** 2163–2181. MR4528496 https://doi.org/10.1080/01621459.2021.1909599

FIKSEL, J., GILBERT, B., WILSON, E., KALTER, H., KANTE, A., AKUM, A., BLAU, D., BASSAT, Q., MACICAME, I. et al. (2023). Correcting for verbal autopsy misclassification bias in cause-specific mortality estimates. *Am. J. Trop. Med. Hyg.* **108** 66–77. https://doi.org/10.4269/ajtmh.22-0318

FLAXMAN, A. D., VAHDATPOUR, A., GREEN, S., JAMES, S. L. and MURRAY, C. J. (2011). Random forests for verbal autopsy analysis: Multisite validation study using clinical diagnostic gold standards. *Popul. Health Metr.* **9** 29. https://doi.org/10.1186/1478-7954-9-29

GARENNE, M. (2014). Prospects for automated diagnosis of verbal autopsies. *BMC Med.* **12** 1–5.

GILBERT, B., FIKSEL, J., WILSON, E., KALTER, H., KANTE, A., AKUM, A., BLAU, D., BASSAT, Q., MACICAME, I. et al. (2023). Multi-cause calibration of verbal autopsy–based cause-specific mortality estimates of children and neonates in Mozambique. *Am. J. Trop. Med. Hyg.* **108** 78–89. https://doi.org/10.4269/ajtmh.22-0319

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

GUSTAFSON, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology*: *Impacts and Bayesian Adjustments*. *Interdisciplinary Statistics*. CRC Press/CRC, Boca Raton, FL. MR2005104

HOOTEN, M. B., JOHNSON, D. S. and BROST, B. M. (2021). Making recursive Bayesian inference accessible. *Amer. Statist.* **75** 185–194. MR4256123 https://doi.org/10.1080/00031305.2019.1665584

JAMES, S. L., FLAXMAN, A. D. and MURRAY, C. J. (2011). Performance of the tariff method: Validation of a simple additive algorithm for analysis of verbal autopsies. *Popul. Health Metr.* **9** 31. https://doi.org/10.1186/1478-7954-9-31

KALTER, H. D., ROUBANATOU, A.-M., KOFFI, A. and BLACK, R. E. (2015). Direct estimates of national neonatal and child cause-specific mortality proportions in Niger by expert algorithm and physician-coded analysis of verbal autopsy interviews. *J. Glob. Health* **5** 010415. https://doi.org/10.7189/jogh.05.010415

KING, G. and LU, Y. (2008). Verbal autopsy methods with multiple causes of death. *Statist. Sci.* **23** 78–91. MR2523943 https://doi.org/10.1214/07-STS247

KOOPMAN, B., KARIMI, S., NGUYEN, A., MCGUIRE, R., MUSCATELLO, D., KEMP, M., TRURAN, D., ZHANG, M. and THACKWAY, S. (2015). Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC Med. Inform. Decis. Mak.* **15** 53. https://doi.org/10.1186/s12911-015-0174-2

LI, Z. R., THOMAS, J., CHOI, E., MCCORMICK, T. H. and CLARK, S. J. (2023). The openVA toolkit for verbal autopsies. *R J.*.

LI, Z. R., THOMAS, J., MCCORMICK, T. H. and CLARK, S. J. (2024a). OpenVA: Automated method for verbal autopsy. R package version 1.1.2.

LI, Z. R., WU, Z., CHEN, I. and CLARK, S. J. (2024b). Bayesian nested latent class models for cause-of-death assignment using verbal autopsies across multiple domains. *Ann. Appl. Stat.* **18** 1137–1159. MR4728659 https://doi.org/10.1214/23-aoas1826

MACICAME, I., KANTE, A., WILSON, E. et al. (2023). Countrywide mortality surveillance for action in Mozambique: Results from a national sample-based vital statistics system for mortality and cause of death. *Am. J. Trop. Med. Hyg.* **108** 5–16. https://doi.org/10.4269/ajtmh.22-0367

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

MENENDEZ, C., CASTILLO, P., MARTÍNEZ, M. J., JORDAO, D., LOVANE, L., ISMAIL, M. R., CARRILHO, C., LORENZONI, C., FERNANDES, F. et al. (2017). Validity of a minimally invasive autopsy for cause of death determination in stillborn babies and neonates in Mozambique: An observational study. *PLoS Med.* **14** e1002318.

MIASNIKOF, P., GIANNAKEAS, V., GOMES, M., ALEKSANDROWICZ, L., SHESTOPALOFF, A. Y., ALAM, D., TOLLMAN, S., SAMARIKHALAJ, A. and JHA, P. (2015). Naive Bayes classifiers for verbal autopsies: Comparison to physician-based classification for 21,000 child and adult deaths. *BMC Med.* **13** 286. https://doi.org/10.1186/s12916-015-0521-2

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

MINSKY, M. (1961). Steps toward artificial intelligence. *Proc. IRE* **49** 8–30. MR0134428

NICHOLS, E. K., BYASS, P., CHANDRAMOHAN, D., CLARK, S. J., FLAXMAN, A. D., JAKOB, R., LEITAO, J., MAIRE, N., RAO, C. et al. (2018). The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0. *PLoS Med.* **15** 1–9. https://doi.org/10.1371/journal.pmed.1002486

PERIN, J., MULICK, A., YEUNG, D., VILLAVICENCIO, F., LOPEZ, G., STRONG, K. L., PRIETO-MERINO, D., COUSENS, S., BLACK, R. E. et al. (2022). Global, regional, and national causes of under-5 mortality in 2000–19: An updated systematic analysis with implications for the sustainable development goals. *Lancet Child Adolesc. Health* **6** 106–115.

POLSON, N. G. and SCOTT, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics* 9 501–538. Oxford Univ. Press, Oxford. With discussions by Bertrand Clark, C. Severinski, Merlise A. Clyde, Robert L. Wolpert, Jim E. Griffin, Philiip J. Brown, Chris Hans, Luis R. Pericchi, Christian P. Robert and Julyan Arbel. MR3204017 https://doi.org/10.1093/acprof:oso/9780199694587.003.0017

PRAMANIK, S., ZEGER, S., BLAU, D. and DATTA, A. (2025). Supplement to "Modeling structure and country-specific heterogeneity in misclassification matrices of verbal autopsy-based cause of death classifiers." https://doi.org/10.1214/24-AOAS2006SUPPA, https://doi.org/10.1214/24-AOAS2006SUPPB

SALZBERG, N. T., SIVALOGAN, K., BASSAT, Q., TAYLOR, A. W., ADEDINI, S., EL ARIFEEN, S., ASSEFA, N., BLAU, D. M., CHAWANA, R. et al. (2019). Mortality surveillance methods to identify and characterize deaths in child health and mortality prevention surveillance network sites. *Clin. Infect. Dis.* **69** S262–S273. https://doi.org/10.1093/cid/ciz599

SETEL, P. W., SANKOH, O., RAO, C., VELKOFF, V. A., MATHERS, C., GONGHUAN, Y., HEMED, Y., JHA, P. and LOPEZ, A. D. (2005). Sample registration of vital events with verbal autopsy: A renewed commitment to measuring and monitoring vital statistics. *Bull. World Health Organ.* **83** 611–617.

SINNOTT, J. A., DAI, W., LIAO, K. P., SHAW, S. Y., ANANTHAKRISHNAN, A. N., GAINER, V. S., KARLSON, E. W., CHURCHILL, S., SZOLOVITS, P. et al. (2014). Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Hum. Genet.* **133** 1369–1382.

SOLEMAN, N., CHANDRAMOHAN, D. and SHIBUYA, K. (2006). Verbal autopsy: Current practices and challenges. *Bull. World Health Organ.* **84** 239–245. https://doi.org/10.2471/blt.05.027003

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

TAYLOR, I., KAPLAN, A. and BETANCOURT, B. (2023). Generative filtering for recursive Bayesian inference with streaming data. arXiv preprint arXiv:2309.14271.

WANG, B., YAO, Y., VISWANATH, B., ZHENG, H. and ZHAO, B. Y. (2018). With great training comes great vulnerability: Practical attacks against transfer learning. In 27*th USENIX Security Symposium* (*USENIX Security* 18) 1281–1297. USENIX Association, Baltimore, MD.

WANG, H., ABAJOBIR, A. A., ABATE, K. H., ABBAFATI, C., ABBAS, K. M., ABD-ALLAH, F., ABERA, S. F., ABRAHA, H. N., ABU-RADDAD, L. J. et al. (2017). Global, regional, and national under-5 mortality, adult mortality, age-specific mortality, and life expectancy, 1970–2016: A systematic analysis for the global burden of disease study 2016. *Lancet* **390** 1084–1150.

WU, Z., LI, Z. R., CHEN, I. and LI, M. (2024). Tree-informed Bayesian multi-source domain adaptation: Cross-population probabilistic cause-of-death assignment using verbal autopsy. *Biostatistics* **25** 1233–1253. MR4808883 https://doi.org/10.1093/biostatistics/kxae005

YANG, Y. and DUNSON, D. B. (2013). Sequential Markov chain Monte Carlo. arXiv preprint arXiv:1308.3861.

YI, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification*: *Strategy*, *Method and Application*. *Springer Series in Statistics*. Springer, New York. With a foreword by Raymond J. Carroll. MR3676914 https://doi.org/10.1007/978-1-4939-6640-0

ZHUANG, F., QI, Z., DUAN, K., XI, D., ZHU, Y., ZHU, H., XIONG, H. and HE, Q. (2021). A comprehensive survey on transfer learning. *Proc*. *IEEE* **109** 43–76. https://doi.org/10.1109/JPROC.2020.3004555

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1240–1269
https://doi.org/10.1214/24-AOAS2009

A PRIVACY-PRESERVED AND HIGH-UTILITY SYNTHESIS STRATEGY FOR RISK-BASED STRATIFIED SUBGROUPS OF THE CANADIAN SCLERODERMA PATIENT REGISTRY DATA

BY BEI JIANG[1,a], ADRIAN E. RAFTERY[2,b], RUSSELL J. STEELE[3,c] AND NAISYIN WANG[4,d]

[1]*Department of Mathematical and Statistical Sciences, University of Alberta,* [a]*bei1@ualberta.ca*
[2]*Department of Statistics and Department of Sociology, University of Washington,* [b]*raftery@uw.edu*
[3]*Department of Mathematics and Statistics, McGill University,* [c]*russell.steele@mcgill.ca*
[4]*Department of Statistics, University of Michigan,* [d]*nwangaa@umich.edu*

Responsible data sharing anchors research reproducibility and promotes the integrity of scientific research. Motivated by Canadian Scleroderma Research Group (CSRG) patient registry data, we present a risk-based method to produce privacy-preserved and high-utility synthetic datasets, which also simultaneously imputes missing data of mixed continuous and categorical types in the original dataset. This method divides all individuals into different subgroups, based on their reidentification risks, and provides tailored synthesis strategies targeted for each risk subgroup, through the associated tuning mechanisms. Under our setting, our risk-based method reduced the number of patients at risk from 198 to four, among the 691 CSRG patients who have no missing values in any of the quasi-identifying variables, while preserving all correct inferential conclusions in the target analysis. The 95% confidence intervals (CIs) have 92.6% overlap, on average, with the CIs constructed using the unperturbed imputation-completed datasets. These findings suggest that our risk-based method makes it possible to release complete synthetic datasets for research reproducibility while ensuring that the reidentification risks are acceptably low. In contrast, the existing one-size-fits-all synthesis strategies that do not take account of different risk levels can lead to unnecessary information loss and possibly incorrect scientific conclusions.

REFERENCES

ATKINSON, R. D. and CORY, N. (2021). Cross-border data policy: Opportunities and challenges. In *Consensus or Conflict? China and Globalization in the* 21*st Century* 217–232.

BARTH-JONES, D. (2012). The "re-identification" of governor William Weld's medical information: A critical re-examination of health data identification risks and privacy protections, then and now. *Then Now*.

CHEN, R. J., LU, M. Y., CHEN, T. Y., WILLIAMSON, D. F. and MAHMOOD, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nat. Biomed. Eng.* **5** 493–497.

DALENIUS, T. (1977). Towards a methodology for statistical disclosure control. *Statistik Tidskrift* **15** 429–444.

DE VRIES-BOUWSTRA, J. K., ALLANORE, Y., MATUCCI-CERINIC, M. and BALBIR-GURMAN, A. (2020). Worldwide expert agreement on updated recommendations for the treatment of systemic sclerosis. *J. Rheumatol.* **47** 249–254. https://doi.org/10.3899/jrheum.181173

DOMINGO-FERRER, J., SÁNCHEZ, D. and BLANCO-JUSTICIA, A. (2021). The limits of differential privacy (and its misuse in data release and machine learning). *Commun. ACM* **64** 33–35.

DOVE, E. S. (2018). The EU general data protection regulation: Implications for international scientific research in the digital era. *J. Law Med. Ethics* **46** 1013–1030.

DRECHSLER, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Lecture Notes in Statistics* **201**. Springer, New York. MR2809912 https://doi.org/10.1007/978-1-4614-0326-5

DRECHSLER, J. (2018). Some clarifications regarding fully synthetic data. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD* 2018, *Valencia, Spain, September* 26–28, 2018, *Proceedings* 109–121. Springer.

Key words and phrases. Mixed continuous and discrete missing data, data augmentation, copula model, reidentification risk, multiple imputation.

DRECHSLER, J. and REITER, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Comput*. *Statist*. *Data Anal*. **55** 3232–3243. MR2825406 https://doi.org/10.1016/j.csda.2011.06.006

DUNCAN, G. and LAMBERT, D. (1989). The risk of disclosure for microdata. *J. Bus*. *Econom*. *Statist*. **7** 207–217.

DUNCAN, G. T. and LAMBERT, D. (1986). Disclosure-limited data dissemination. *J. Amer. Statist*. *Assoc*. **81** 10–18.

DUSETZINA, S. B., TYREE, S., MEYER, A.-M., MEYER, A., GREEN, L. and CARPENTER, W. R. (2014). Linking data for health services research: A framework and instructional guide.

DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*. *Lecture Notes in Computer Science* **3876** 265–284. Springer, Berlin. MR2241676 https://doi.org/10.1007/11681878_14

EL EMAM, K. (2013). *Guide to the De-identification of Personal Health Information*. CRC Press.

FELDMAN, J. and KOWAL, D. R. (2022). Bayesian data synthesis and the utility-risk trade-off for mixed epidemiological data. *Ann*. *Appl*. *Stat*. **16** 2577–2602. MR4489224 https://doi.org/10.1214/22-aoas1604

FELLEGI, I. P. (1972). On the question of statistical confidentiality. *J. Amer. Statist*. *Assoc*. **67** 7–18.

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

GRUND, S., LÜDTKE, O. and ROBITZSCH, A. (2024). Using synthetic data to improve the reproducibility of statistical results in psychological research. *Psychol*. *Methods* **29** 789–806. https://doi.org/10.1037/met0000526

HAWALA, S. (2008). Producing partially synthetic data to avoid disclosure. In *Proceedings of the Joint Statistical Meetings*. American Statistical Association, Alexandria, VA.

HOTZ, V. J., BOLLINGER, C. R., KOMAROVA, T., MANSKI, C. F., MOFFITT, R. A., NEKIPELOV, D., SOJOURNER, A. and SPENCER, B. D. (2022). Balancing data privacy and usability in the federal statistical system. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **119** e2104906119.

HU, J., SAVITSKY, T. and WILLIAMS, M. (2020). Risk-weighted data synthesizers for microdata dissemination. *Chance* **33** 29–36.

HU, J., SAVITSKY, T. D. and WILLIAMS, M. R. (2021). Risk-efficient Bayesian pseudo posterior data synthesis for privacy protection. *J. Surv. Stat*. *Methodol*.. smab013. https://doi.org/10.1093/jssam/smab013

HUDSON, M., STEELE, R., LU, Y., THOMBS, B. D., CANADIAN SCLERODERMA RESEARCH GROUP and BARON, M. (2009). Work disability in systemic sclerosis. *J. Rheumatol*. **36** 2481–2486. https://doi.org/10.3899/jrheum.081237

JIANG, B., RAFTERY, A. E., STEELE, R. J. and WANG, N. (2022). Balancing inferential integrity and disclosure risk via model targeted masking and multiple imputation. *J. Amer. Statist*. *Assoc*. **117** 52–66. MR4399067 https://doi.org/10.1080/01621459.2021.1909597

JIANG, B., RAFTERY, A. E., STEELE, R. J. and WANG, N. (2025). Supplement to "A privacy-preserved and high-utility synthesis strategy for risk-based stratified subgroups of the Canadian scleroderma patient registry data." https://doi.org/10.1214/24-AOAS2009SUPPA, https://doi.org/10.1214/24-AOAS2009SUPPB

LITTLE, R. J. (1993). Statistical analysis of masked data. *J. Off. Stat*. **9** 407–426.

LITTLE, R. J. A., LIU, F. and RAGHUNATHAN, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. *Wiley Ser. Probab. Stat*. 141–152. Wiley, Chichester. MR2138251 https://doi.org/10.1002/0470090456.ch13

LIU, J. S. and WU, Y. N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist*. *Assoc*. **94** 1264–1274. MR1731488 https://doi.org/10.2307/2669940

LO, B., COETZEE, T., DEMETS, D. et al. (2015). *Sharing Clinical Trial Data*: *Maximizing Benefits*, *Minimizing Risk*. National Academies Press. ISBN 978-0-309-31629-3.

LOONG, B., ZASLAVSKY, A. M., HE, Y. and HARRINGTON, D. P. (2013). Discloure control using partially synthetic data for large-scale health surveys, with applications to CanCORS. *Stat*. *Med*. **32** 4139–4161. MR3118346 https://doi.org/10.1002/sim.5841

MORRISROE, K., HUDSON, M., BARON, M., DE VRIES-BOUWSTRA, J., CARREIRA, P. E., WUTTGE, D. M., WANG, M., FRECH, T. M., STEVENS, W. et al. (2018). Determinants of health-related quality of life in a multinational systemic sclerosis inception cohort. *Clin*. *Exp*. *Rheumatol*. **36** 53–60.

MYERS, R. H., MONTGOMERY, D. C. and ANDERSON-COOK, C. M. (2016). *Response Surface Methodology*: *Process and Product Optimization Using Designed Experiments*, 4th ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3497873

NOWOK, B., RAAB, G. M. and DIBBEN, C. (2016). synthpop: Bespoke creation of synthetic data in R. *J. Stat*. *Softw*. **74** 1–26.

OHM, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Rev*. **57** 1701.

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

RAGHUNATHAN, T. E. (2021). Synthetic data. *Annu. Rev. Stat. Appl.* **8** 129–140. MR4243543 https://doi.org/10.1146/annurev-statistics-040720-031848

REITER, J. P. (2005). Estimating risks of identification disclosure in microdata. *J. Amer. Statist. Assoc.* **100** 1103–1112. MR2236926 https://doi.org/10.1198/016214505000000619

ROCHER, L., HENDRICKX, J. M. and DE MONTJOYE, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* **10** 3069.

RUBIN, D. B. (1993). Statistical disclosure limitation. *J. Off. Stat.* **9** 461–468.

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Classics Library*. Wiley Interscience, Hoboken, NJ. Reprint of the 1987 edition [John Wiley & Sons, Inc., New York; MR899519]. MR2117498

RUBIO-RIVAS, M. (2021). Nationwide and international registries on scleroderma. Past, present, and future. *Span. J. Med.* **1** 142–153.

SNOKE, J., RAAB, G. M., NOWOK, B., DIBBEN, C. and SLAVKOVIC, A. (2018). General and specific utility measures for synthetic data. *J. Roy. Statist. Soc. Ser. A* **181** 663–688. MR3807503 https://doi.org/10.1111/rssa.12358

SWEENEY, L. (2002). k-anonymity: A model for protecting privacy. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems* **10** 557–570. 5, Aggregation and security assessment for inference control in statistical databases. MR1948199 https://doi.org/10.1142/S0218488502001648

THOMBS, B. D., HUDSON, M., BASSEL, M., TAILLEFER, S. S., BARON, M. and GGROUP, C. S. R. (2009). Sociodemographic, disease, and symptom correlates of fatigue in systemic sclerosis: Evidence from a sample of 659 Canadian scleroderma research group registry patients. *Arthritis Care Res.* **61** 966–973.

WILLENBORG, L. and DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. *Lecture Notes in Statistics* **155**. Springer, New York. MR1866909 https://doi.org/10.1007/978-1-4613-0121-9

WINKLER, W. E. (1997). Views on the production and use of confidential microdata. U.S. Census Bureau Research Report No. RR97/01.

WINKLER, W. E. (2007). Examples of easy-to-implement, widely used masking methods for which analytic properties are not justified. U.S. Census Bureau Research Report No. RRS2007-21.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1270–1291
https://doi.org/10.1214/25-AOAS2017

FEDERATED LEARNING OF ROBUST INDIVIDUALIZED DECISION RULES WITH APPLICATION TO HETEROGENEOUS MULTIHOSPITAL SEPSIS POPULATION

BY XINLEI CHEN[1,a], VICTOR B. TALISA[2,d], XIAOQING TAN[1,b], ZHENGLING QI[3,g], JASON N. KENNEDY[2,e], CHUNG-CHOU H. CHANG[4,h], CHRISTOPHER W. SEYMOUR[2,f] AND LU TANG[1,c]

[1]*Department of Biostatistics and Health Data Science, University of Pittsburgh,* [a]*xic120@pitt.edu,* [b]*xit31@pitt.edu,* [c]*lutang@pitt.edu*

[2]*Department of Critical Care Medicine, University of Pittsburgh,* [d]*vit13@pitt.edu,* [e]*jnk28@pitt.edu,* [f]*seymourc@pitt.edu*

[3]*School of Business, George Washington University,* [g]*qizhengling@email.gwu.edu*

[4]*Department of Medicine, University of Pittsburgh,* [h]*changj@pitt.edu*

Sepsis is a life-threatening condition affecting millions of individuals in the U.S. each year. The complexity of sepsis clinical management makes individualized treatment approaches desirable. The University of Pittsburgh Medical Center (UPMC) has collected electronic health records data of sepsis patients from multiple hospitals. The goal of this study is to derive individualized decision rules (IDRs) that could be safely applied to and uniformly improve decision-making across hospitals in the UPMC Health System by only using a subset of hospitals for training. Traditional approaches assume that data are sampled from a single population of interest. With multiple hospitals that vary in patient populations, treatments, and provider teams, an IDR that is successful in one hospital may not be as effective in another, and the performance achieved by a globally optimal IDR may vary greatly across hospitals, preventing it from being safely applied to unseen hospitals. To address these challenges as well as the practical restriction of data sharing across hospitals, we introduce a new objective function and a federated learning algorithm for learning IDRs that are robust to distributional uncertainty from heterogeneous data. The proposed framework uses a conditional maximin objective to enhance individual outcomes across hospitals, ensuring robustness against hospital-level variations. Compared to the traditional approach, the proposed method enhances the survival rate by 10 percentage points among patients who may experience extreme adverse outcomes across hospitals. Additionally, it increases the overall survival rate by two to three percentage points when the learned IDR is applied to unseen hospital populations.

REFERENCES

ANGUS, D. C., LINDE-ZWIRBLE, W. T., LIDICKER, J., CLERMONT, G., CARCILLO, J. and PINSKY, M. R. (2001). Epidemiology of severe sepsis in the United States: Analysis of incidence, outcome, and associated costs of care. *Crit. Care Med.* **29** 1303–1310. https://doi.org/10.1097/00003246-200107000-00002

ANGUS, D. C. and VAN DER POLL, T. (2013). Severe sepsis and septic shock. *N. Engl. J. Med.* **369** 840–851. https://doi.org/10.1056/NEJMra1208623

AZARM-DAIGLE, M., KUZIEMSKY, C. and PEYTON, L. (2015). A review of cross organizational healthcare data sharing. *Proc. Comput. Sci.* **63** 425–432.

BAI, X., YU, W., JI, W., LIN, Z., TAN, S., DUAN, K., DONG, Y., XU, L. and LI, N. (2014). Early versus delayed administration of norepinephrine in patients with septic shock. *J. Crit. Care* **18** 1–8.

BLANCHARD, P., EL MHAMDI, E. M., GUERRAOUI, R. and STAINER, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. *Adv. Neural Inf. Process. Syst.* **30**.

Key words and phrases. Causal inference, conditional average treatment effect, data integration, distributionally robust learning, decentralized data.

BRISIMI, T. S., CHEN, R., MELA, T., OLSHEVSKY, A., PASCHALIDIS, I. C. and SHI, W. (2018). Federated learning of predictive models from federated Electronic Health Records. *Int. J. Med. Inform.* **112** 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007

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

CARRANZA, A. G. and ATHEY, S. (2023). Federated offline policy learning with heterogeneous observational data. arXiv preprint. Available at arXiv:2305.12407.

CHEN, X., TALISA, V. B., TAN, X., QI, Z., KENNEDY, J. N., CHANG, C.-C., SEYMOUR, C. W. and TANG, L. (2025). Supplement to "Federated Learning of Robust Individualized Decision Rules with Application to Heterogeneous Multihospital Sepsis Population." https://doi.org/10.1214/25-AOAS2017SUPPA, https://doi.org/10.1214/25-AOAS2017SUPPB

CLIFTON, J. and LABER, E. (2020). Q-learning: Theory and applications. *Annu. Rev. Stat. Appl.* **7** 279–301. MR4104194 https://doi.org/10.1146/annurev-statistics-031219-041220

EVANS, L., RHODES, A., ALHAZZANI, W., ANTONELLI, M., COOPERSMITH, C. M., FRENCH, C., MACHADO, F. R., MCINTYRE, L., OSTERMANN, M. et al. (2021). Surviving sepsis campaign: International guidelines for management of sepsis and septic shock 2021. *Intensive Care Med.* **47** 1181–1247.

FALLAH, A., MOKHTARI, A. and OZDAGLAR, A. (2020a). On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International Conference on Artificial Intelligence and Statistics* 1082–1092. PMLR.

FALLAH, A., MOKHTARI, A. and OZDAGLAR, A. (2020b). Personalized federated learning: a meta-learning approach. arXiv preprint. Available at arXiv:2002.07948.

FANG, E. X., WANG, Z. and WANG, L. (2023). Fairness-oriented learning for optimal individualized treatment rules. *J. Amer. Statist. Assoc.* **118** 1733–1746. MR4646602 https://doi.org/10.1080/01621459.2021.2008402

GAO, J., XU, Y., BARREIRO-GOMEZ, J., NDONG, M., SMYRNAKIS, M. and TEMBINE, H. (2018). Distributionally robust optimization. In *Algorithms-Examplesm*, *Vol.* 1.

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

GUHA, N., TALWALKAR, A. and SMITH, V. (2019). One-shot federated learning. arXiv preprint. Available at arXiv:1902.11175.

HERNÁN, M. A. and ROBINS, J. M. (2010). *Causal Inference*: *What If*.

HUANG, L., SHEA, A. L., QIAN, H., MASURKAR, A., DENG, H. and LIU, D. (2019). Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *J. Biomed. Inform.* **99** 103291.

INVESTIGATORS AND THE ANZICS CLINICAL TRIALS GROUP (2014). Goal-directed resuscitation for patients with early septic shock. *N. Engl. J. Med.* **371** 1496–1506.

JIANG, Y., KONEČNÝ, J., RUSH, K. and KANNAN, S. (2019). Improving federated learning personalization via model agnostic meta learning. arXiv preprint. Available at arXiv:1909.12488.

KENNEDY, J. N. (2021). Towards a Learning Health System: Using Reinforcement Learning to Optimize Treatment Decisions in Sepsis Patients.

KIM, Y., SUN, J., YU, H. and JIANG, X. (2017). Federated tensor factorization for computational phenotyping. In *Proceedings of the* 23*rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 887–895.

KOMOROWSKI, M., CELI, L. A., BADAWI, O., GORDON, A. C. and FAISAL, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* **24** 1716–1720. https://doi.org/10.1038/s41591-018-0213-5

LAHA, N., SONABEND-W, A., MUKHERJEE, R. and CAI, T. (2024). Finding the optimal dynamic treatment regimes using smooth Fisher consistent surrogate loss. *Ann. Statist.* **52** 679–707. MR4744192 https://doi.org/10.1214/24-aos2363

LEE, J., SUN, J., WANG, F., WANG, S., JUN, C.-H., JIANG, X. et al. (2018). Privacy-preserving patient similarity learning in a federated environment: Development and analysis. *JMIR Med. Inform.* **6** e7744.

LI, T., HU, S., BEIRAMI, A. and SMITH, V. (2021). Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning* 6357–6368. PMLR.

LI, T., SANJABI, M., BEIRAMI, A. and SMITH, V. (2019). Fair resource allocation in federated learning. arXiv preprint. Available at arXiv:1905.10497.

MARIK, P. E., LINDE-ZWIRBLE, W. T., BITTNER, E. A., SAHATJIAN, J. and HANSELL, D. (2017). Fluid administration in severe sepsis and septic shock, patterns and outcomes: An analysis of a large national database. *Intensive Care Med.* **43** 625–632. https://doi.org/10.1007/s00134-016-4675-y

MCMAHAN, B., MOORE, E., RAMAGE, D., HAMPSON, S. and ARCAS, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics* 1273–1282. PMLR.

MO, W., QI, Z. and LIU, Y. (2021). Learning optimal distributionally robust individualized treatment rules. *J. Amer. Statist. Assoc.* **116** 659–674. MR4270012 https://doi.org/10.1080/01621459.2020.1796359

MOUNCEY, P. R., OSBORN, T. M., POWER, G. S., HARRISON, D. A., SADIQUE, M. Z., GRIEVE, R. D., JAHAN, R., HARVEY, S. E., BELL, D. et al. (2015). Trial of early, goal-directed resuscitation for septic shock. *N. Engl. J. Med.* **372** 1301–1311.

MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 331–355. MR1983752 https://doi.org/10.1111/1467-9868.00389

MURPHY, S. A. (2005). A generalization error for Q-learning. *J. Mach. Learn. Res.* **6** 1073–1097. MR2249849

PANTANOWITZ, A. and MARWALA, T. (2009). Missing data imputation through the use of the random forest algorithm. In *Advances in Computational Intelligence* 53–62. Springer, Berlin.

PFOHL, S. R., DAI, A. M. and HELLER, K. (2019). Federated and differentially private learning for electronic health records. arXiv preprint. Available at arXiv:1911.05861.

PILLUTLA, K., KAKADE, S. M. and HARCHAOUI, Z. (2022). Robust aggregation for federated learning. *IEEE Trans. Signal Process.* **70** 1142–1154. MR4396314 https://doi.org/10.1109/tsp.2022.3153135

NHLBI PREVENTION AND EARLY TREATMENT OF ACUTE LUNG INJURY CLINICAL TRIALS NETWORK (2023). Early restrictive or liberal fluid management for sepsis-induced hypotension. *N. Engl. J. Med.* **388** 499–510. https://doi.org/10.1056/NEJMoa2212663

QI, Z., CUI, Y., LIU, Y. and PANG, J.-S. (2019). Estimation of individualized decision rules based on an optimized covariate-dependent equivalent of random outcomes. *SIAM J. Optim.* **29** 2337–2362. MR4008648 https://doi.org/10.1137/18M1190975

QI, Z. and LIU, Y. (2018). D-learning to estimate optimal individual treatment rules. *Electron. J. Stat.* **12** 3601–3638. MR3870507 https://doi.org/10.1214/18-ejs1480

QI, Z., PANG, J.-S. and LIU, Y. (2023). On robustness of individualized decision rules. *J. Amer. Statist. Assoc.* **118** 2143–2157. MR4646632 https://doi.org/10.1080/01621459.2022.2038180

RAGHU, A., KOMOROWSKI, M., AHMED, I., CELI, L., SZOLOVITS, P. and GHASSEMI, M. (2017). Deep reinforcement learning for sepsis treatment. arXiv preprint. Available at arXiv:1711.09602.

RAHIMIAN, H. and MEHROTRA, S. (2019). Distributionally robust optimization: a review. arXiv preprint. Available at arXiv:1908.05659.

RHEE, C., DANTES, R., EPSTEIN, L., MURPHY, D. J., SEYMOUR, C. W., IWASHYNA, T. J., KADRI, S. S., ANGUS, D. C., DANNER, R. L. et al. (2017). Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014. *JAMA* **318** 1241–1249.

RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071 https://doi.org/10.1198/016214504000001880

SEYMOUR, C. W., KENNEDY, J. N., WANG, S., CHANG, C.-C. H., ELLIOTT, C. F., XU, Z., BERRY, S., CLERMONT, G., COOPER, G. et al. (2019). Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *J. Amer. Med. Assoc.* **321** 2003–2017.

SHAH, F., TALISA, V., CHANG, C.-C., TRIANTAFYLLOU, S., TANG, L., MAYR, F., HIGGINS, A., PEAKE, S., MOUNCEY, P. et al. (2024). Heterogeneity in the effect of early goal-directed therapy for septic shock: A secondary analysis of two multicenter international trials. *Crit. Care Med.*

SINGER, M., DEUTSCHMAN, C. S., SEYMOUR, C. W., SHANKAR-HARI, M., ANNANE, D., BAUER, M., BELLOMO, R., BERNARD, G. R., CHICHE, J.-D. et al. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *J. Amer. Med. Assoc.* **315** 801–810.

SUN, Y. and WANG, L. (2021). Stochastic tree search for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.* **116** 421–432. MR4227704 https://doi.org/10.1080/01621459.2020.1819294

SUN, Z., KAIROUZ, P., SURESH, A. T. and MCMAHAN, H. B. (2019). Can you really backdoor federated learning? arXiv preprint. Available at arXiv:1911.07963.

TAN, X., QI, Z., SEYMOUR, C. and TANG, L. (2022). RISE: Robust individualized decision learning with sensitive variables. *Adv. Neural Inf. Process. Syst.* **35** 19484–19498.

TANG, L. and SONG, P. X. K. (2016). Fused lasso approach in regression coefficients clustering—learning parameter heterogeneity in data integration. *J. Mach. Learn. Res.* **17** Paper No. 113, 23. MR3543519

TAO, Y., WANG, L. and ALMIRALL, D. (2018). Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. *Ann. Appl. Stat.* **12** 1914–1938. MR3852703 https://doi.org/10.1214/18-AOAS1137

TIAN, Y. and FENG, Y. (2023). Transfer learning under high-dimensional generalized linear models. *J. Amer. Statist. Assoc.* **118** 2684–2697. MR4681613 https://doi.org/10.1080/01621459.2022.2071278

WAECHTER, J., KUMAR, A., LAPINSKY, S. E., MARSHALL, J., DODEK, P., ARABI, Y., PARRILLO, J. E., DELLINGER, R. P., GARLAND, A. et al. (2014). Interaction between fluids and vasoactive agents on mortality in septic shock: A multicenter, observational study. *Crit. Care Med.* **42** 2158–2168.

WANG, L., ZHOU, Y., SONG, R. and SHERWOOD, B. (2018). Quantile-optimal treatment regimes. *J. Amer. Statist. Assoc.* **113** 1243–1254. MR3862354 https://doi.org/10.1080/01621459.2017.1330204

WANG, Y., FU, H. and ZENG, D. (2018). Learning optimal personalized treatment rules in consideration of benefit and risk: With an application to treating type 2 diabetes patients with insulin therapies. *J. Amer. Statist. Assoc.* **113** 1–13. MR3803435 https://doi.org/10.1080/01621459.2017.1303386

WANG, Z., WANG, J., ZHOU, Q., LI, B. and LI, H. (2022). Sample-efficient reinforcement learning via conservative model-based actor-critic. In *Proceedings of the AAAI Conference on Artificial Intelligence* 8612–8620.

WIESEMANN, W., KUHN, D. and SIM, M. (2014). Distributionally robust convex optimization. *Oper. Res.* **62** 1358–1376. MR3294550 https://doi.org/10.1287/opre.2014.1314

ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68** 1010–1018. MR3040007 https://doi.org/10.1111/j.1541-0420.2012.01763.x

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

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1292–1313
https://doi.org/10.1214/24-AOAS1996

CECNN: COPULA-ENHANCED CONVOLUTIONAL NEURAL NETWORKS IN JOINT PREDICTION OF REFRACTION ERROR AND AXIAL LENGTH BASED ON ULTRA-WIDEFIELD FUNDUS IMAGES

BY CHONG ZHONG[1,a], YANG LI[2,c], DANJUAN YANG[3,e], MEIYAN LI[3,f], XINGTAO ZHOU[3,g], BO FU[2,d], CATHERINE C. LIU[1,b] AND A. H. WELSH[4,h]

[1]*Department of Data Science and Artificial Intelligence, Hong Kong Polytechnic University,* [a]*chzhong@polyu.edu.hk,* [b]*macliu@polyu.edu.hk*

[2]*School of Data Science, Fudan University,* [c]*18110980006@fudan.edu.cn,* [d]*fu@fudan.edu.cn*

[3]*Eye Institute and Department of Ophthalmology, Eye & ENT Hospital, Fudan University,* [e]*docdanjuanyang@126.com,* [f]*limeiyan0406073@126.com,* [g]*doctzhouxingtao@126.com*

[4]*College of Business and Economics, Australian National University,* [h]*Alan.Welsh@anu.edu.au*

The ultra-widefield (UWF) fundus image is an attractive 3D biomarker in AI-aided myopia screening because it provides much richer myopia-related information. Though axial length (AL) has been acknowledged to be highly related to the two key targets of myopia screening, spherical equivalence (SE) measurement and high myopia diagnosis, its prediction based on the UWF fundus image is rarely considered. To save the high expense and time costs of measuring SE and AL, we propose the Copula-enhanced Convolutional Neural Network (CeCNN), a one-stop UWF-based ophthalmic AI framework to jointly predict SE, AL, and myopia status. The CeCNN formulates a multiresponse regression that relates multiple dependent discrete-continuous responses and the image covariate, where the nonlinearity of the association is modeled by a backbone CNN. To thoroughly describe the dependence structure among the responses, we model and incorporate the conditional dependence among responses in a CNN through a new copula-likelihood loss. We provide statistical interpretations of the conditional dependence among responses and reveal that such dependence is beyond the dependence explained by the image covariate. We heuristically justify that the proposed loss can enhance the estimation efficiency of the CNN weights. We apply the CeCNN to the UWF dataset collected by us and demonstrate that the CeCNN sharply enhances the predictive capability of various backbone CNNs. Our study supports the ophthalmology view that, besides SE, AL is an important measure of myopia.

REFERENCES

AVERY, R. B. (1977). Error components and seemingly unrelated regressions. *Econometrica* **45** 199–209. MR0451543 https://doi.org/10.2307/1913296

CEN, L.-P., JI, J., LIN, J.-W., JU, S.-T., LIN, H.-J., LI, T.-P., WANG, Y., YANG, J.-F., LIU, Y.-F. et al. (2021). Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat. Commun.* **12** 4828.

CHEN, Z.-M., WEI, X.-S., WANG, P. and GUO, Y. (2019). Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5177–5186.

DAI, B., SHEN, X. and PAN, W. (2024). Significance tests of feature relevance for a black-box learner. *IEEE Trans. Neural Netw. Learn. Syst.* **35** 1898–1911. MR4710268 https://doi.org/10.1109/tnnls.2022.3185742

DE'ATH, G. (2002). Multivariate regression trees: A new technique for modeling species–environment relationships. *Ecology* **83** 1105–1117.

DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., DEHGHANI, M., MINDERER, M., HEIGOLD, G. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Key words and phrases. Copula, convolutional neural network, multitask learning, myopia, ultra-widefield fundus image, 3D medical image object.

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

HAARMAN, A. E., ENTHOVEN, C. A., TIDEMAN, J. W. L., TEDJA, M. S., VERHOEVEN, V. J. and KLAVER, C. C. (2020). The complications of myopia: A review and meta-analysis. *Investig*. *Ophthalmol*. *Vis*. *Sci*. **61** 49–49.

HAN, S., POOL, J., TRAN, J. and DALLY, W. (2015). Learning both weights and connections for efficient neural network. *Adv*. *Neural Inf*. *Process*. *Syst*. **28**.

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.

HUANG, G., LIU, Z., VAN DER MAATEN, L. and WEINBERGER, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4700–4708.

IWASE, A., ARAIE, M., TOMIDOKORO, A., YAMAMOTO, T., SHIMIZU, H., KITAZAWA, Y., GROUP, T. S. et al. (2006). Prevalence and causes of low vision and blindness in a Japanese adult population: The Tajimi study. *Ophthalmology* **113** 1354–1362.

JAIN, P. and TEWARI, A. (2015). Alternating minimization for regression problems with vector-valued outputs. *Adv*. *Neural Inf*. *Process*. *Syst*. **28**.

KENDALL, A., GAL, Y. and CIPOLLA, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 7482–7491.

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

KIM, K. M., HEO, T.-Y., KIM, A., KIM, J., HAN, K. J., YUN, J. and MIN, J. K. (2021). Development of a fundus image-based deep learning diagnostic tool for various retinal diseases. *J*. *Personalized Med*. **11** 321.

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. Preprint. Available at arXiv:1412.6980.

KOBAYASHI, K., OHNO-MATSUI, K., KOJIMA, A., SHIMADA, N., YASUZUMI, K., YOSHIDA, T., FUTAGAMI, S., TOKORO, T. and MOCHIZUKI, M. (2005). Fundus characteristics of high myopia in children. *Jpn*. *J*. *Ophthalmol*. **49** 306–311.

LAI, Q., ZHOU, J., GAN, Y., VONG, C.-M. and CHEN, C. P. (2024). Single-stage broad multi-instance multi-label learning (BMIML) with diverse inter-correlations and its application to medical image classification. *IEEE Trans*. *Signal Process*. **8** 828–839.

LECUN, Y., BENGIO, Y. and HINTON, G. (2015). Deep learning. *Nature* **521** 436–444.

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

LI, Y., HUANG, Q., ZHONG, C., YANG, D., LI, M., WELSH, A., LIU, A., FU, B., LIU, C. C. et al. (2024). OUCopula: Bi-channel multi-label copula-enhanced adapter-based CNN for myopia screening based on OU-UWF images. In *IJCAI*.

LI, Z., GUO, C., LIN, D., NIE, D., ZHU, Y., CHEN, C., ZHAO, L., WANG, J., ZHANG, X. et al. (2021). Deep learning for automated glaucomatous optic neuropathy detection from ultra-widefield fundus images. *Br*. *J*. *Ophthalmol*. **105** 1548–1554.

LIAN, C., LIU, M., WANG, L. and SHEN, D. (2022). Multi-task weakly-supervised attention network for dementia status estimation with structural MRI. *IEEE Trans*. *Neural Netw*. *Learn*. *Syst*. **33** 4056–4068.

LIN, X., ZHEN, H.-L., LI, Z., ZHANG, Q.-F. and KWONG, S. (2019). Pareto multi-task learning. *Adv*. *Neural Inf*. *Process*. *Syst*. **32**.

LIU, M., ZHANG, J., ADELI, E. and SHEN, D. (2019). Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. *IEEE Trans*. *Biomed*. *Eng*. **66** 1195–1206.

LOH, W.-Y. and ZHENG, W. (2013). Regression trees for longitudinal and multiresponse data. *Ann*. *Appl*. *Stat*. **7** 495–522. MR3086428 https://doi.org/10.1214/12-AOAS596

MENG, W., BUTTERWORTH, J., MALECAZE, F. and CALVAS, P. (2011). Axial length of myopia: A review of current research. *Ophthalmologica* **225** 127–134.

MIDENA, E., MARCHIONE, G., DI GIORGIO, S., ROTONDI, G., LONGHIN, E., FRIZZIERO, L., PILOTTO, E., PARROZZANI, R. and MIDENA, G. (2022). Ultra-wide-field fundus photography compared to ophthalmoscopy in diagnosing and classifying major retinal diseases. *Sci*. *Rep*. **12** 19287.

MUTTI, D. O., HAYES, J. R., MITCHELL, G. L., JONES, L. A., MOESCHBERGER, M. L., COTTER, S. A., KLEINSTEIN, R. N., MANNY, R. E., TWELKER, J. D. et al. (2007). Refractive error, axial length, and relative peripheral refractive error before and after the onset of myopia. *Investig*. *Ophthalmol*. *Vis*. *Sci*. **48** 2510–2519.

OH, R., LEE, E. K., BAE, K., PARK, U. C., YU, H. G. and YOON, C. K. (2023). Deep learning-based prediction of axial length using ultra-widefield fundus photography. Korean journal of ophthalmology. *Korean J*. *Ophthalmol*. **37** 95–104.

ÖZGÜR, A. and NAR, F. (2020). Effect of dropout layer on classical regression problems. In 2020 28*th Signal Processing and Communications Applications Conference* (*SIU*) 1–4. IEEE.

PANAGIOTELIS, A., CZADO, C. and JOE, H. (2012). Pair copula constructions for multivariate discrete data. *J. Amer. Statist. Assoc.* **107** 1063–1072. MR3010894 https://doi.org/10.1080/01621459.2012.682850

PRUCHA, I. R. (1984). On the asymptotic efficiency of feasible Aitken estimators for seemingly unrelated regression models with error components. *Econometrica* **52** 203–207. MR0729216 https://doi.org/10.2307/1911468

RAHMAN, R., OTRIDGE, J. and PAL, R. (2017). Integratedmrf: Random forest-based framework for integrating prediction from different data types. *Bioinformatics* **33** 1407–1410.

RASKUTTI, G., YUAN, M. and CHEN, H. (2019). Convex regularization for high-dimensional multiresponse tensor regression. *Ann. Statist.* **47** 1554–1584. MR3911122 https://doi.org/10.1214/18-AOS1725

SELVARAJU, R. R., COGSWELL, M., DAS, A., VEDANTAM, R., PARIKH, D. and BATRA, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* 618–626.

SKLAR, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **8** 229–231. MR0125600

SONG, L., LIU, J., QIAN, B., SUN, M., YANG, K., SUN, M. and ABBAS, S. (2018). A deep multi-modal CNN for multi-instance multi-label image classification. *IEEE Trans. Image Process.* **27** 6025–6038. MR3852320 https://doi.org/10.1109/TIP.2018.2864920

SONG, P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications. Springer Series in Statistics*. Springer, New York. MR2377853

SONG, P. X.-K., LI, M. and YUAN, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics* **65** 60–68. MR2665846 https://doi.org/10.1111/j.1541-0420.2008.01058.x

SUN, K., HE, M., XU, Y., WU, Q., HE, Z., LI, W., LIU, H. and PI, X. (2022). Multi-label classification of fundus images with graph convolutional network and lightgbm. *Comput. Biol. Med.* **149** 105909.

TIDEMAN, J. W. L., SNABEL, M. C., TEDJA, M. S., VAN RIJN, G. A., WONG, K. T., KUIJPERS, R. W., VINGERLING, J. R., HOFMAN, A., BUITENDIJK, G. H. et al. (2016). Association of axial length with risk of uncorrectable visual impairment for europeans with myopia. *JAMA Ophthalmol.* **134** 1355–1363.

YANG, L., FREES, E. W. and ZHANG, Z. (2020). Nonparametric estimation of copula regression models with discrete outcomes. *J. Amer. Statist. Assoc.* **115** 707–720. MR4107674 https://doi.org/10.1080/01621459.2018.1546586

YANG, S., XIAO, W., ZHANG, M., GUO, S. ZHAO, J. and SHEN, F. (2022). Image data augmentation for deep learning: A survey. Preprint. Available at arXiv:2204.08610.

ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57** 348–368. MR0139235

ZHANG, S., CHEN, Y., LI, Z., WANG, W., XUAN, M., ZHANG, J., HU, Y., CHEN, Y., XIAO, O. et al. (2024). Axial elongation trajectories in Chinese children and adults with high myopia. *JAMA Ophthalmol.* **142** 87–94.

ZHONG, C., LI, Y., YANG, D., LI, M., ZHOU, X., FU, B., LIU, C. C. and WELSH, A. H. (2025). Supplement to "CeCNN: Copula-enhanced convolutional neural networks in joint prediction of refraction error and axial length based on ultra-widefield fundus images." https://doi.org/10.1214/24-AOAS1996SUPPA, https://doi.org/10.1214/24-AOAS1996SUPPB

ZOU, C., KE, Y. and ZHANG, W. (2022). Estimation of low rank high-dimensional multivariate linear models for multi-response data. *J. Amer. Statist. Assoc.* **117** 693–703. MR4436306 https://doi.org/10.1080/01621459.2020.1799813

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1314–1331
https://doi.org/10.1214/25-AOAS2013

A DEEP NEURAL NETWORK TWO-PART MODEL AND FEATURE IMPORTANCE TEST FOR SEMICONTINUOUS DATA

BY BAIMING ZOU[1,a], XINLEI MI[2,e], SHIYU WAN[1,b], DI WU[1,c], JAMES G. XENAKIS[3,f], JIANHUA HU[4,g] AND FEI ZOU[1,d]

[1]*Department of Biostatistics, University of North Carolina at Chapel Hill,* [a]*bzou@email.unc.edu,* [b]*shiyu31@unc.edu,* [c]*did@email.unc.edu,* [d]*feizou@email.unc.edu*

[2]*Gilead Science,* [e]*xlmi@email.unc.edu*

[3]*Harvard University,* [f]*jxenakis@fas.harvard.edu*

[4]*Columbia University,* [g]*jh3992@cumc.columbia.edu*

Semicontinuous data frequently arise in clinical practice. For example, while many surgical patients still suffer from varying degrees of acute postoperative pain (POP) sometime after surgery (i.e., POP score > 0), others experience none (i.e., POP score $= 0$), indicating the existence of two distinct data processes at play. Existing parametric or semiparametric two-part modeling methods for this type of semicontinuous data can fail to appropriately model the two underlying data processes, as such methods rely heavily on (generalized) linear additive assumptions. However, many factors may interact to jointly influence the experience of POP nonadditively and nonlinearly. Motivated by this challenge and inspired by the flexibility of deep neural networks (DNN) to accurately approximate complex functions universally, we derive a DNN-based two-part model by adapting the conventional DNN methods with two additional components: a bootstrapping procedure along with a filtering algorithm to boost the stability of the conventional DNN, an approach we denote as sDNN. To improve the interpretability and transparency of sDNN, we further derive a feature importance testing procedure to identify important features associated with the outcome measurements of the two data processes, denoting this approach fsDNN. We show that fsDNN not only offers a statistical inference procedure for each feature under complex association but also that using the identified features can further improve the predictive performance of sDNN. The proposed sDNN- and fsDNN-based two-part models are applied to the analysis of real data from a POP study in which application they clearly demonstrate advantages over the existing parametric and semiparametric two-part models. Further, we conduct extensive numerical studies and draw comparisons with other machine learning methods to demonstrate that sDNN and fsDNN consistently outperform the existing two-part models and frequently used machine learning methods regardless of the data complexity. An R package implementing the proposed methods has been developed and is available in the Supplementary Material (Zou et al., (2025)) and is also deposited on GitHub (https://github.com/BZou-lab/fsDNN).

REFERENCES

ALTMANN, A., TOLOŞI, L., SANDER, O. and LENGAUER, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics* **26** 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

AMIT, Y. and GEMAN, D. (1997). Shape quantization and recognition with randomized trees. *Neural Comput*. **9** 1545–1588.

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

BAYMAN, E. O., OLESON, J. J. and RABBITTS, J. A. (2021). AAAPT: Assessment of the acute pain trajectory. *Pain Med*. **22** 533–547. https://doi.org/10.1093/pm/pnaa440

Key words and phrases. Complex function, nonlinearity, nonadditive, permutation, profile, testable machine learning model.

BECK, N. and JACKMAN, S. (1998). Beyond linearity by default: Generalized additive models. *Amer. J. Polit. Sci.* **42** 596–627.

BELOTTI, F. et al. (2015). twopm: Two-part models. *Stata J.* **15** 3–20.

BENGIO, Y. (2009). Learning deep architectures for AI. *Found. Trends Mach. Learn.* **2** 1–127.

BENGIO, Y. (2012). Practical recommendations for gradient-based training of deep architectures. Available at arXiv:1206.5533.

BOT, A. G. J., BEKKERS, S., ARNSTEIN, P. M., SMITH, R. M. and RING, D. (2014). Opioid use after fracture surgery correlates with pain intensity and satisfaction with pain relief. *Clin. Orthop. Relat. Res.* **472** 2542–2549. https://doi.org/10.1007/s11999-014-3660-4

BREIMAN, L. (1996). Bagging predictors. *Mach. Learn.* **24** 123–140.

BYRD, R. H., CHIN, G. M., NOCEDAL, J. and WU, Y. (2012). Sample size selection in optimization methods for machine learning. *Math. Program.* **134** 127–155. MR2947555 https://doi.org/10.1007/s10107-012-0572-5

CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265

CHE, N., HOLLY, V. H., BRIAN, S. C., SHAHA, N. D. and ROZALINA, G. M. (2019). Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c. *J. Biomed. Inform.* **89** 56–67.

CHOU, R. et al. (2016). Guideline on the management of postoperative pain. *J. Pain* **17** 131–157.

CRAVEN, M. and SHAVLIK, J. W. (1996). Extracting tree-structured representations of trained networks. In *Advances in Neural Information Processing Systems* 24–30.

CYBENKO, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2** 303–314. MR1015670 https://doi.org/10.1007/BF02551274

DUAN, N., MANNING, W., MORRIS, C. and NEWHOUSE, J. (1983). A comparison of alternative models for the demand for medical care. *J. Bus. Econom. Statist.* **1** 115–126.

FARAGO', A. and LUGOSI, G. (1993). Strong universal consistency of neural network classifiers. *IEEE Trans. Inf. Theory* **39** 1146–1151.

FAREWELL, V. T., LONG, D. L., TOM, B. D. M., YIU, S. and SU, L. (2017). Two-part and related regression models for longitudinal data. *Annu. Rev. Stat. Appl.* **4** 283–315. https://doi.org/10.1146/annurev-statistics-060116-054131

GAN, T. J. (2017). Poorly controlled postoperative pain: Prevalence, consequences, and prevention. *J. Pain Res.* **10** 2287–2298. https://doi.org/10.2147/JPR.S144066

GLOROT, X. and BENGIO, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* **9** 249–256.

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

HANSEN, L. and SALAMON, P. (1990). Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* **12** 993–1001.

HARRELL, F., LEE, K. and MARK, D. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15** 361–367.

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

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

HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.* **2** 359–366.

KEHLET, H., JENSEN, T. S. and WOOLF, C. J. (2006). Persistent postsurgical pain: Risk factors and prevention. *Lancet* **367** 1618–1625. https://doi.org/10.1016/S0140-6736(06)68700-X

KINGA, D. and ADAM, J. B. (2015). A method for stochastic optimization. In *International Conference on Learning Representations* (*ICLR*) **5**.

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

LOVICH-SAPOLA, J., SMITH, C. E. and BRANDT, C. P. (2015). Postoperative pain control. *Surg. Clin. North Amer.* **95** 301–318. https://doi.org/10.1016/j.suc.2014.10.002

LU, Y., FAN, Y., LV, J. and NOBLE, W. S. (2018). DeepPINK: Reproducible feature selection in deep neural networks. In *Advances in Neural Information Processing Systems* 8676–8686.

LU, Z., PU, H., WANG, F., HU, Z. and WANG, L. (2017). The expressive power of neural networks: A view from the width. *Neural Inf. Process. Syst.* 6231–6239.

LUNDBERG, S. M. and LEE, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 4765–4774.

MEI, S., MONTANARI, A. and NGUYEN, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA* **115** E7665–E7671. MR3845070 https://doi.org/10.1073/pnas.1806579115

MI, X., ZOU, B., ZOU, F. and HU, J. (2021). Permutation-based identification of important biomarkers for complex diseases via machine learning models. *Nat. Commun.* **12** 3008.

MI, X., ZOU, F. and ZHU, R. (2019). Bagging and deep learning in optimal individualized treatment rules. *Biometrics* **75** 674–684. MR3999189 https://doi.org/10.1111/biom.12990

MIHAYLOVA, B. et al. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Econ.* **20** 897–916.

MIN, Y. and AGRESTI, A. (2002). Modeling nonnegative data with clumping at zero: A survey. *J. Iran. Stat. Soc.* (*JIRSS*) **1** 7–33.

NAIR, A., VELAGAPUDI, M., LANG, J., BEHARA, L., VENIGANDLA, R., VELAGAPUDI, N. et al. (2020). Machine learning approach to predict postoperative opioid requirements in ambulatory surgery patients. *PLoS ONE* **15** e0236833.

OLSEN, M. K. and SCHAFER, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *J. Amer. Statist. Assoc.* **96** 730–745. MR1946438 https://doi.org/10.1198/016214501753168389

PUTIN, E., MAMOSHINA, P., ALIPER, A., KORZINKIN, M., MOSKALEV, A., KOLOSOV, A., OSTROVSKIY, A., CANTOR, C., VIJG, J. et al. (2016). Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging* (*Albany NY*) **8** 1021–1033.

RIBEIRO, M. T., SINGH, S. and GUESTRIN, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144. ACM, New York.

SANTOSA, F. and SYMES, W. W. (1986). Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Statist. Comput.* **7** 1307–1330. MR0857796 https://doi.org/10.1137/0907087

SAUZET, O., RAZUM, O., WIDERA, T. and BRZOSKA, P. (2019). Two-part models and quantile regression for the analysis of survey data with a spike. The example of satisfaction with health care. *Front. Public Health* **7** 1–7.

SHAPLEY, L. S. (1953). A value for n-person games. In *Contributions to the Theory of Games*, *Vol.* 2. *Ann. of Math. Stud.*, *No.* 28 307–317. Princeton Univ. Press, Princeton, NJ. MR0053477

SMALL, C. and LAYCOCK, H. (2020). Acute postoperative pain management. *Br. J. Surg.* **107** e70–e80. https://doi.org/10.1002/bjs.11477

SONODA, S. and MURATA, N. (2017). Neural network with unbounded activation functions is universal approximator. *Appl. Comput. Harmon. Anal.* **43** 233–268. MR3668038 https://doi.org/10.1016/j.acha.2015.12.005

TAYLOR, J. (1995). Semi-parametric estimation in failure time mixture distributions. *Biometrics* **51** 814–817.

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

TIGHE, P. et al. (2015a). Teaching a machine to feel postoperative pain: Combining high-dimensional clinical data with machine learning algorithms to forecast acute postoperative pain. *Pain Med.* **16** 1386–1401.

TIGHE, P. J., KING, C. D., ZOU, B. and FILLINGIM, R. B. (2016). Time to onset of Sustained Postoperative Pain Relief (SuPPR): Evaluation of a new systems-level metric for acute pain management. *Clin. J. Pain* **32** 371–379. https://doi.org/10.1097/AJP.0000000000000285

TIGHE, P. J., LE-WENDLING, L. T., PATEL, A., ZOU, B. and FILLINGIM, R. B. (2015b). Clinically derived early postoperative pain trajectories differ by age, sex, and type of surgery. *Pain* **156** 609–617. https://doi.org/10.1097/01.j.pain.0000460352.07836.0d

TRIKHA, A. and SINGH, P. M. (2013). Predicting post-operative pain: Still a long way to go! *J. Anaesthesiol.-Clin. Pharmacol.* **29** 433–434.

TU, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* **49** 1225–1231. https://doi.org/10.1016/s0895-4356(96)00002-9

YANG, M., HARTLEY, R., LEUNG, A. et al. (2019). Preoperative predictors of poor acute postoperative pain control: A systematic review and meta-analysis. *BMJ Open* **9** e025091.

ZHENG, H., SCHNABEL, A., YAHIAOUI-DOKTOR, M., MEISSNER, W., VAN AKEN, H., ZAHN, P. et al. (2017). Age and preoperative pain are major confounders for sex differences in postoperative pain outcome: A prospective database analysis. *PLoS ONE* **12** e0178659.

ZHOU, X.-H. and LIANG, H. (2006). Semi-parametric single-index two-part regression models. *Comput. Statist. Data Anal.* **50** 1378–1390. MR2224377 https://doi.org/10.1016/j.csda.2004.12.001

ZHOU, Z.-H., WU, J. and TANG, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence* **137** 239–263. MR1906477 https://doi.org/10.1016/S0004-3702(02)00190-X

ZOU, B., MI, X., WAN, S., WU, D., XENAKIS, J. G, HU, J. and ZOU, F. (2025). Supplement to "A deep neural network two-part model and feature importance test for semicontinuous data." https://doi.org/10.1214/25-AOAS2013SUPPA, https://doi.org/10.1214/25-AOAS2013SUPPB

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1332–1361
https://doi.org/10.1214/24-AOAS2007

BAYESIAN DATA AUGMENTATION FOR RECURRENT EVENTS UNDER INTERMITTENT ASSESSMENT IN OVERLAPPING INTERVALS WITH APPLICATIONS TO EMR DATA

BY XIN LIU[a] AND PATRICK M. SCHNELL[b]

Division of Biostatistics, College of Public Health, The Ohio State University, [a]*liu.7302@osu.edu,* [b]*schnell.31@osu.edu*

Electronic medical records (EMR) data contain rich information that can facilitate health-related studies but is collected primarily for purposes other than research. For recurrent events, EMR data often do not record event times or counts but only contain intermittently assessed and censored observations (i.e., upper and/or lower bounds for counts in a time interval) at uncontrolled times. This can result in noncontiguous or overlapping assessment intervals with censored event counts. Existing methods for analyzing intermittently assessed recurrent events assume disjoint assessment intervals with known counts (interval count data) due to a focus on prospective studies with controlled assessment times. We propose a Bayesian data augmentation method to analyze the complicated assessments in EMR data for recurrent events. Within a Gibbs sampler, event times are imputed by generating sets of event times from nonhomogeneous Poisson processes and rejecting proposed sets that are incompatible with constraints imposed by assessment data. Based on the independent increments property of Poisson processes, we implement three techniques to speed up this rejection sampling imputation method for large EMR datasets: independent sampling by partitioning, truncated generation, and sequential sampling. In a simulation study, we show our method accurately estimates parameters of log-linear Poisson process intensities. Although the proposed method can be applied generally to EMR data of recurrent events, our study is specifically motivated by identifying risk factors for falls due to cancer treatment and its supportive medications. We used the proposed method to analyze an EMR dataset comprising 5501 patients treated for breast cancer. Our analysis provides evidence supporting associations between certain risk factors (including classes of medications) and risk of falls.

REFERENCES

ALEXANDER, B. H., RIVARA, F. P. and WOLF, M. E. (1992). The cost and frequency of hospitalization for fall-related injuries in older adults. *Amer. J. Publ. Health* **82** 1020–1023. https://doi.org/10.2105/ajph.82.7.1020

BEA, J. W., ZHAO, Q., CAULEY, J. A., LACROIX, A. Z., BASSFORD, T., LEWIS, C. E., JACKSON, R. D., TYLAVSKY, F. A. and CHEN, Z. (2011). Effect of hormone therapy on lean body mass, falls, and fractures: Six-year results from the women's health initiative hormone trials. *Menopause* **18** 44–52. https://doi.org/10.1097/gme.0b013e3181e3aab1

CHEN, B. E., COOK, R. J., LAWLESS, J. F. and ZHAN, M. (2005). Statistical methods for multivariate interval-censored recurrent events. *Stat. Med.* **24** 671–691. MR2134533 https://doi.org/10.1002/sim.1936

ÇINLAR, E. (1975). *Introduction to Stochastic Processes*. Prentice Hall, Englewood Cliffs, NJ. MR0380912

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

COX, D. R. and LEWIS, P. A. W. (1966). *The Statistical Analysis of Series of Events*. Methuen, London. MR0199942

DE JONG, M. R., VAN DER ELST, M. and HARTHOLT, K. A. (2013). Drug-related falls in older patients: Implicated drugs, consequences, and possible prevention strategies. *Ther. Adv. Drug Saf.* **4** 147–154. https://doi.org/10.1177/2042098613486829

Key words and phrases. Bayesian, data augmentation, Poisson process, recurrent event, electronic medical records, electronic health record, intermittent assessment.

DUNSON, D. B., HOLLOMAN, C., CALDER, C. and GUNN, L. H. (2004). Bayesian modeling of multiple lesion onset and growth from interval-censored data. *Biometrics* **60** 676–683. MR2089443 https://doi.org/10.1111/j.0006-341X.2004.00217.x

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

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

GEYER, C. J. (2007). Lower-Truncated Poisson and Negative Binomial Distributions. Univ. Minnesota, MN. Accessed Feb 2024.

HELSKE, J. (2021). ramcmc: Robust Adaptive Metropolis Algorithm R package version 0.1.2.

HERNÁN, M. A., BRUMBACK, B. and ROBINS, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* **11** 561–570. https://doi.org/10.1097/00001648-200009000-00012

HUANG, M. H., BLACKWOOD, J., GODOSHIAN, M. and PFALZER, L. (2019). Predictors of falls in older survivors of breast and prostate cancer: A retrospective cohort study of surveillance, epidemiology and end results—medicare health outcomes survey linkage. *J. Geriatr. Oncol.* **10** 89–97. https://doi.org/10.1016/j.jgo.2018.04.009

JANVIN, M., YOUNG, J. G., RYALEN, P. C. and STENSRUD, M. J. (2024). Causal inference with recurrent and competing events. *Lifetime Data Anal.* **30** 59–118. MR4684786 https://doi.org/10.1007/s10985-023-09594-8

KASS, R. E., CARLIN, B. P., GELMAN, A. and NEAL, R. M. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *Amer. Statist.* **52** 93–100. MR1628427 https://doi.org/10.2307/2685466

LAWLESS, J. F. and ZHAN, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Canad. J. Statist.* **26** 549–565. https://doi.org/10.2307/3315717

LIU, X. and SCHNELL, P. M. (2025). Supplement to "Bayesian data augmentation for recurrent events under intermittent assessment in overlapping intervals with applications to EMR data." https://doi.org/10.1214/24-AOAS2007SUPPA, https://doi.org/10.1214/24-AOAS2007SUPPB

LORENTZON, M., JOHANSSON, H., HARVEY, N. C., LIU, E., VANDENPUT, L., CRANDALL, C. J., CAULEY, J. A., LEBOFF, M. S., MCCLOSKEY, E. V. et al. (2022). Menopausal hormone therapy reduces the risk of fracture regardless of falls risk or baseline FRAX probability-results from the women's health initiative hormone therapy trials. *Osteoporos. Int.* **33** 2297–2305. https://doi.org/10.1007/s00198-022-06483-y

MERRICK, J. R. and SOYER, R. (2017). Semiparametric Bayesian optimal replacement policies: Application to railroad tracks. *Appl. Stoch. Models Bus. Ind.* **33** 445–460. MR3718843 https://doi.org/10.1002/asmb.2210

MOHILE, S. G., FAN, L., REEVE, E., JEAN-PIERRE, P., MUSTIAN, K., PEPPONE, L., JANELSINS, M., MORROW, G., HALL, W. et al. (2011). Association of cancer with geriatric syndromes in older medicare beneficiaries. *J. Clin. Oncol.* **29** 1458–1464. https://doi.org/10.1200/JCO.2010.31.6695

PANDYA, C., MAGNUSON, A., DALE, W., LOWENSTEIN, L., FUNG, C. and MOHILE, S. G. (2016). Association of falls with health-related quality of life (HRQOL) in older cancer survivors: A population based study. *J. Geriatr. Oncol.* **7** 201–210. https://doi.org/10.1016/j.jgo.2016.01.007

PASUPATHY, R. (2011a). Generating homogeneous Poisson processes. In *Wiley Encyclopedia of Operations Research and Management Science Eorms*0355 Wiley, Hoboken, NJ, USA. https://doi.org/10.1002/9780470400531.eorms0355

PASUPATHY, R. (2011b). Generating nonhomogeneous Poisson processes. In *Wiley Encyclopedia of Operations Research and Management Science Eorms*0356. Wiley, Hoboken, NJ, USA. https://doi.org/10.1002/9780470400531.eorms0356

RANDELL, K. M., HONKANEN, R. J., KOMULAINEN, M. H., TUPPURAINEN, M. T., KRÖGER, H. and SAARIKOSKI, S. (2001). Hormone replacement therapy and risk of falling in early postmenopausal women—a population-based study. *Clin. Endocrinol.* **54** 769–774. https://doi.org/10.1046/j.1365-2265.2001.01300.x

RESNICK, S. (1992). *Adventures in Stochastic Processes*. Birkhäuser, Boston, MA. MR1181423

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581

SATTAR, S., ALIBHAI, S. M. H., SPOELSTRA, S. L., FAZELZAD, R. and PUTS, M. T. E. (2016). In falls in older adults with cancer: A systematic review of prevalence, injurious falls, and impact on cancer treatment. *Support. Care Cancer* **24** 4459–4469. https://doi.org/10.1007/s00520-016-3342-8

SATTAR, S., ALIBHAI, S. M. H., SPOELSTRA, S. L. and PUTS, M. T. E. (2019). The assessment, management, and reporting of falls, and the impact of falls on cancer treatment in community-dwelling older patients receiving cancer treatment: Results from a mixed-methods study. *J. Geriatr. Oncol.* **10** 98–104. https://doi.org/10.1016/j.jgo.2018.08.006

SHEN, H. and COOK, R. J. (2015). Analysis of interval-censored recurrent event processes subject to resolution. *Biom. J.* **57** 725–742. MR3394807 https://doi.org/10.1002/bimj.201400162

SPOELSTRA, S. L., GIVEN, B. A., SCHUTTE, D. L., SIKORSKII, A., YOU, M. and GIVEN, C. W. (2013). Do older adults with cancer fall more often? A comparative analysis of falls in those with and without cancer. *Oncol. Nurs. Forum* **40** E69–E78. https://doi.org/10.1188/13.ONF.E69-E78

STANISWALIS, J. G., THALL, P. F. and SALCH, J. (1997). Semiparametric regression analysis for recurrent event interval counts. *Biometrics* **53** 1334–1353. MR1614386 https://doi.org/10.2307/2533501

STEVENS, J. A., BALLESTEROS, M. F., MACK, K. A., RUDD, R. A., DECARO, E. and ADLER, G. (2012). Gender differences in seeking care for falls in the aged medicare population. *Am. J. Prev. Med.* **43** 59–62. https://doi.org/10.1016/j.amepre.2012.03.008

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

THALL, P. F. (1988). Mixed Poisson likelihood regression models for longitudinal interval count data. *Biometrics* **44** 197–209. MR0931635 https://doi.org/10.2307/2531907

THALL, P. F. and LACHIN, J. M. (1988). Analysis of recurrent events: Nonparametric methods for random-interval count data. *J. Amer. Statist. Assoc.* **83** 339–347.

TOFTHAGEN, C., OVERCASH, J. and KIP, K. (2012). Falls in persons with chemotherapy-induced peripheral neuropathy. *Support. Care Cancer* **20** 583–589. https://doi.org/10.1007/s00520-011-1127-7

TOOMEY, A. and FRIEDMAN, L. (2014). Mortality in cancer patients after a fall-related injury: The impact of cancer spread and type. *Injury* **45** 1710–1716. https://doi.org/10.1016/j.injury.2014.03.008

VIHOLA, M. (2012). Robust adaptive Metropolis algorithm with coerced acceptance rate. *Stat. Comput.* **22** 997–1008. MR2950080 https://doi.org/10.1007/s11222-011-9269-5

WHO (2021). Falls. Available at https://www.who.int/news-room/fact-sheets/detail/falls.

WOOLCOTT, J. C., RICHARDSON, K. J., WIENS, M. O., PATEL, B., MARIN, J., KHAN, K. M. and MARRA, C. A. (2009). Meta-analysis of the impact of 9 medication classes on falls in elderly persons. *Arch. Intern. Med.* **169** 1952–1960. https://doi.org/10.1001/archinternmed.2009.357

ZHANG, X., SUN, M., LIU, S., LEUNG, C. H., PANG, L., POPAT, U. R., CHAMPLIN, R., HOLMES, H. M., VALERO, V. et al. (2018). Risk factors for falls in older patients with cancer. *BMJ Support. Palliat. Care* **8** 34–37. https://doi.org/10.1136/bmjspcare-2017-001388

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1362–1381
https://doi.org/10.1214/25-AOAS2018

A FLEXIBLE AND PARSIMONIOUS MODELLING STRATEGY FOR CLUSTERED DATA ANALYSIS

BY TAO HUANG[1,a], YOUQUAN PEI[2,c], JINHONG YOU[1,b] AND WENYANG ZHANG[3,d]

[1]*School of Statistics and Management, Shanghai University of Finance and Economics,* [a]*huang.tao@mail.sufe.edu.cn,* [b]*johnyou07@163.com*

[2]*School of Economics, Shandong University,* [c]*ypeistat@sdu.edu.cn*

[3]*Faculty of Business Administration, University of Macau,* [d]*wenyangzhang@um.edu.mo*

The development of an appropriate statistical modelling strategy is of paramount importance for the successful analysis of data. The trade-off between flexibility and parsimony is of vital importance in statistical modelling. In the context of clustered data analysis, it is essential to account for the inherent heterogeneity between clusters while simultaneously ensuring parsimony to mitigate the potential for complexity and to preserve the homogeneity within clusters. The objective of this paper is to propose a flexible and parsimonious modelling strategy for clustered data analysis. The strategy strikes an optimal balance between flexibility and parsimony, effectively accounting for both heterogeneity and homogeneity among the clusters, which often possess significant practical implications. In particular, we apply this modelling strategy to analyse data related to the spread of COVID-19 in China, examining how factors such as human mobility and temperature can capture dynamic and nonlinear patterns in the data. We develop an estimation procedure for the unknown parameters, establish the asymptotic properties of the estimators, and conduct simulation studies to evaluate the performance of our method. Additionally, the real data analysis illustrates the practical application of our approach in understanding regional differences in epidemic spread, with implications for public health policy. This flexible framework also offers insights for transfer learning scenarios beyond clustered data analysis.

REFERENCES

BAHADORI, M. T., KALE, D., FAN, Y. and LIU, Y. (2015). Functional subspace clustering with application to time series. In *ICML*'15: *Proceedings of the* 32*nd International Conference on Machine Learning* **37**. 228–237.

BERTOZZI, A. L., FRANCO, E., MOHLER, G., SHORT, M. B. and SLEDGE, D. (2020). The challenges of modeling and forecasting the spread of COVID-19. *Proc. Natl. Acad. Sci. USA* **117** 16732–16738. MR4242725 https://doi.org/10.1073/pnas.2006520117

CHEN, Z., FAN, J. and LI, R. (2018). Error variance estimation in ultrahigh-dimensional additive models. *J. Amer. Statist. Assoc.* **113** 315–327. MR3803467 https://doi.org/10.1080/01621459.2016.1251440

CHENG, M.-Y., HONDA, T. and ZHANG, J.-T. (2016). Forward variable selection for sparse ultra-high dimensional varying coefficient models. *J. Amer. Statist. Assoc.* **111** 1209–1221. MR3561943 https://doi.org/10.1080/01621459.2015.1080708

CHU, W., LI, R., LIU, J. and REIMHERR, M. (2020). Feature selection for generalized varying coefficient mixed-effect models with application to obesity GWAS. *Ann. Appl. Stat.* **14** 276–298. MR4085094 https://doi.org/10.1214/19-AOAS1310

FAN, J., FAN, Y., LV, J. and YANG, F. (2022). SIMPLE-RC: Group network inference with non-sharp nulls and neak signals. Available at arXiv:2211.00128.

FENG, S., LI, G., PENG, H. and TONG, T. (2021). Varying-coefficient panel data model with interactive fixed effects. *Statist. Sinica* **31** 935–957. MR4286201 https://doi.org/10.5705/ss.20

HU, L., HUANG, T. and YOU, J. (2019). Estimation and identification of a varying-coefficient additive model for locally stationary processes. *J. Amer. Statist. Assoc.* **114** 1191–1204. MR4011772 https://doi.org/10.1080/01621459.2018.1482753

Key words and phrases. Clustered data analysis, semiparametric models, structure identification.

HUANG, T., PEI, Y., YOU, J. and ZHANG, W. (2025). Supplement to "A flexible and parsimonious modelling strategy for clustered data analysis." https://doi.org/10.1214/25-AOAS2018SUPPA, https://doi.org/10.1214/25-AOAS2018SUPPB

KE, Y., LI, J. and ZHANG, W. (2016). Structure identification in panel data analysis. *Ann. Statist.* **44** 1193–1233. MR3485958 https://doi.org/10.1214/15-AOS1403

KE, Y., LIAN, H. and ZHANG, W. (2022). High-dimensional dynamic covariance matrices with homogeneous structure. *J. Bus. Econom. Statist.* **40** 96–110. MR4356560 https://doi.org/10.1080/07350015.2020.1779079

KE, Z. T., FAN, J. and WU, Y. (2015). Homogeneity pursuit. *J. Amer. Statist. Assoc.* **110** 175–194. MR3338495 https://doi.org/10.1080/01621459.2014.892882

KRAEMER, M. U., YANG, C. H., GUTIERREZ, B., WU, C. H., KLEIN, B., PIGOTT, D. M., OPEN COVID-19 DATA WORKINGGROUP, DU PLESSIS, L., FARIA, N. R. et al. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368** 493–497.

LEI, H., XIA, Y. and QIN, X. (2016). Estimation of semivarying coefficient time series models with ARMA errors. *Ann. Statist.* **44** 1618–1660. MR3519935 https://doi.org/10.1214/15-AOS1430

LI, D., KE, Y. and ZHANG, W. (2015). Model selection and structure specification in ultra-high dimensional generalised semi-varying coefficient models. *Ann. Statist.* **43** 2676–2705. MR3405608 https://doi.org/10.1214/15-AOS1356

LI, G. and LI, W. K. (2008). Testing for threshold moving average with conditional heteroscedasticity. *Statist. Sinica* **18** 647–665. MR2432284

LI, G. and LI, W. K. (2011). Testing a linear time series model against its threshold extension. *Biometrika* **98** 243–250. MR2804225 https://doi.org/10.1093/biomet/asq074

LI, J., ZHANG, W. and KONG, E. (2018). Factor models for asset returns based on transformed factors. *J. Econometrics* **207** 432–448. MR3868190 https://doi.org/10.1016/j.jeconom.2018.09.001

LI, Y. and GENTON, M. G. (2009). Single-index additive vector autoregressive time series models. *Scand. J. Stat.* **36** 369–388. MR2549700 https://doi.org/10.1111/j.1467-9469.2009.00641.x

LIAN, H., QIAO, X. and ZHANG, W. (2021). Homogeneity pursuit in single index models based panel data analysis. *J. Bus. Econom. Statist.* **39** 386–401. MR4235184 https://doi.org/10.1080/07350015.2019.1665531

LIU, J., LI, R. and WU, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *J. Amer. Statist. Assoc.* **109** 266–274. MR3180562 https://doi.org/10.1080/01621459.2013.850086

PEI, Y., HUANG, T., PENG, H. and YOU, J. (2022). Network-based clustering for varying coefficient panel data models. *J. Bus. Econom. Statist.* **40** 578–594. MR4410883 https://doi.org/10.1080/07350015.2020.1841648

REN, Z., KANG, Y., FAN, Y. and LV, J. (2019). Tuning-free heterogeneous inference in massive networks. *J. Amer. Statist. Assoc.* **114** 1908–1925. MR4047309 https://doi.org/10.1080/01621459.2018.1537920

SMITH, T. P., FLAXMAN, S., GALLINAT, A. S., KINOSIAN, S. P., STEMKOVSKI, M., UNWIN, H. J. T., WATSON, O. J., WHITTAKER, C., CATTARINO, L. et al. (2021). Temperature and population density influence SARS-CoV-2 transmission in the absence of nonpharmaceutical interventions. *Proc. Natl. Acad. Sci. USA* **118** e2019284118.

SUN, Y., YAN, H., ZHANG, W. and LU, Z. (2014). A semiparametric spatial dynamic model. *Ann. Statist.* **42** 700–727. MR3210984 https://doi.org/10.1214/13-AOS1201

TANG, F., FENG, Y., CHIHEB, H. and FAN, J. (2021). The interplay of demographic variables and social distancing scores in deep prediction of U.S. COVID-19 cases. *J. Amer. Statist. Assoc.* **116** 492–506. MR4269998 https://doi.org/10.1080/01621459.2021.1901717

VOGT, M. and LINTON, O. (2017). Classification of non-parametric regression functions in longitudinal data models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 5–27. MR3597962 https://doi.org/10.1111/rssb.12155

WANG, T. and XIA, Y. (2015). Whittle likelihood estimation of nonlinear autoregressive models with moving average residuals. *J. Amer. Statist. Assoc.* **110** 1083–1099. MR3420686 https://doi.org/10.1080/01621459.2014.946513

WU, H., FAN, Y. and LV, J. (2020). Statistical insights into deep neural network learning in subspace classification. *Stat* **9** e273. MR4156468 https://doi.org/10.1002/sta4.273

XIAO, D., KE, Y. and LI, R. (2021). Homogeneity structure learning in large-scale panel data with heavy-tailed errors. *J. Mach. Learn. Res.* **22** Paper No. 13. MR4253706

YUAN, M., LIU, R., FENG, Y. and SHANG, Z. (2022). Testing community structure for hypergraphs. *Ann. Statist.* **50** 147–169. MR4382012 https://doi.org/10.1214/21-aos2099

ZHONG, P.-S., LI, R. and SANTO, S. (2019). Homogeneity tests of covariance matrices with high-dimensional longitudinal data. *Biometrika* **106** 619–634. MR3992393 https://doi.org/10.1093/biomet/asz011

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1382–1402
https://doi.org/10.1214/25-AOAS2025

TEMPORAL NETWORK INFLUENCE MODEL WITH APPLICATION TO THE COVID-19 POPULATION FLOW NETWORK

BY DONGXUE ZHANG[1,a], LONG FENG[2,d], YUJIA WU[1,b], WEI LAN[1,c] AND JING ZHOU[3,e]

[1]*School of Statistics and Center of Statistical Research, Southwestern University of Finance and Economics,* [a]*dongxue.win@gmail.com,* [b]*sugars_yujia@foxmail.com,* [c]*lanwei@swufe.edu.cn*

[2]*School of Statistics and Data Science, KLMDASR, LEBPS and LPMC, Nankai University,* [d]*flnankai@nankai.edu.cn*

[3]*Center for Applied Statistics and School of Statistics, Renmin University of China,* [e]*jing.zhou@ruc.edu.cn*

Ever since its outbreak, COVID-19 has been rapidly spreading around the world and has become a significant threat to public health. Past experience has shown that, because of the incubation period, the contemporaneous population flow does not affect the contemporaneous number of cases, but the time-lagged population flow can affect case numbers. Moreover, the population flow networks of different lags can exhibit varying influences on the transmission of COVID-19. However, most existing studies analyze only the static effects of a network, while ignoring its dynamic characteristics. To assess the dynamic effect of a population flow network, we propose a novel temporal network influence (TNIF) model. This model studies the influence of case numbers via three transmission mechanisms of COVID-19: cross-city transmission, within-city transmission and diffusion transmission. Consequently, three dynamic parameters are introduced to measure the three transmission influences, respectively. To estimate these influential parameters, we develop a quasi-maximum likelihood estimator and establish its theoretical properties. Additionally, to examine the changing pattern of the cross-city transmission influential parameter, we provide two max-type test statistics for testing the homogeneity and change point, respectively. We also demonstrate their asymptotic distributions. Furthermore, a Bayesian information criterion is developed to select the order of the within-city and diffusion transmission influential parameters. The effectiveness of the proposed TNIF model is evaluated through simulation studies. Finally, we apply the proposed TNIF model to analyze the COVID-19 data that collected from 283 Chinese cities from January 20, 2020 to April 23, 2020. We found that the population flow network from the previous period and two periods prior had a significant positive impact on the number of cases on the current period, while the network from earlier periods (three or more periods prior) had no significant effect. This finding further implies that the "two-week lockdown" policy implemented at that time was reasonable and could effectively slow down the spread of COVID-19 at a relatively low cost.

REFERENCES

AINSLIE, K. E. C., WALTERS, C. E., FU, H., BHATIA, S., WANG, H., XI, X., BAGUELIN, M., BHATT, S., BOONYASIRI, A. et al. (2020). Evidence of initial success for China exiting COVID-19 social distancing policy after achieving containment. *Wellcome Open Res*. **5** 81. https://doi.org/10.12688/wellcomeopenres.15843.2

ANSELIN, L. and BERA, A. K. (1998). Spatial dependence in linear regression models, with an introduction to spatial econometrics. In *Handbook of Applied Economic Statistics* (A. Ullah and D. E. A. Giles, eds.) 237–289. Dekker, New York.

BICKEL, P. J. and SARKAR, P. (2016). Hypothesis testing for automated community detection in networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **78** 253–273. MR3453655 https://doi.org/10.1111/rssb.12117

BLOMMESTEIN, H. J. (1983). Specification and estimation of spatial econometric models: A discussion of alternative strategies for spatial economic modelling. *Reg. Sci. Urban Econ*. **13** 250–271.

Key words and phrases. COVID-19, dynamic network effect, population flow network, quasi-maximum likelihood estimation, Temporal Network Influence Model.

BLOMMESTEIN, H. J. (1985). Elimination of circular routes in spatial dynamic regression equations. *Reg. Sci. Urban Econ.* **15** 121–130.

BONACCORSI, G., PIERRI, F., CINELLI, M., FLORI, A., GALEAZZI, A., PORCELLI, F., SCHMIDT, A. L., VALENSISE, C. M., SCALA, A. et al. (2020). Economic and social consequences of human mobility restrictions under COVID-19. *Proc. Natl. Acad. Sci. USA* **117** 15530–15535. https://doi.org/10.1073/pnas.2007658117

BRAMOULLÉ, Y., DJEBBARI, H. and FORTIN, B. (2009). Identification of peer effects through social networks. *J. Econometrics* **150** 41–55. MR2525993 https://doi.org/10.1016/j.jeconom.2008.12.021

BRAMOULLÉ, Y., DJEBBARI, H. and FORTIN, B. (2020). Peer effects in networks: A survey. *Ann. Rev. Econ.* **12** 603–629.

CHEN, J., CHEN, W., LIU, E., LUO, J. and SONG, Z. (2022). The economic cost of lockdown in China: Evidence from city-to-city truck flows. Working Paper.

HAN, X., XU, Y., FAN, L., HUANG, Y., XU, M. and GAO, S. (2021). Quantifying COVID-19 importation risk in a dynamic network of domestic cities and international countries. *Proc. Natl. Acad. Sci. USA* **118** e2100201118.

HUANG, J. S. (1984). The autoregressive moving average model for spatial analysis. *Aust. J. Stat.* **26** 169–178. MR0766617 https://doi.org/10.1111/j.1467-842x.1984.tb01231.x

JIA, J. S., LU, X., YUAN, Y., XU, G., JIA, J. and CHRISTAKIS, N. A. (2020). Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* **582** 389–394.

KRAEMER, M. U., YANG, C. H., GUTIERREZ, B., WU, C. H., KLEIN, B., PIGOTT, D. M., DU PLESSIS, L., FARIA, N. R., LI, R. et al. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368** 493–497.

LAUER, S. A., GRANTZ, K. H., BI, Q., JONES, F. K., ZHENG, Q., MEREDITH, H. R., AZMAN, A. S., REICH, N. G. and LESSLER, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann. Intern. Med.* **172** 577–582. https://doi.org/10.7326/M20-0504

LEE, L. and YU, J. (2012). QML estimation of spatial dynamic panel data models with time varying spatial weights matrices. *Spatial Econ. Anal.* **7** 31–74.

LEE, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* **72** 1899–1925. MR2095537 https://doi.org/10.1111/j.1468-0262.2004.00558.x

LEE, S., HA, J., NA, O. and NA, S. (2003). The cusum test for parameter change in time series models. *Scand. J. Stat.* **30** 781–796. MR2155483 https://doi.org/10.1111/1467-9469.00364

LEI, J. (2016). A goodness-of-fit test for stochastic block models. *Ann. Statist.* **44** 401–424. MR3449773 https://doi.org/10.1214/15-AOS1370

MANSKI, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Rev. Econ. Stud.* **60** 531–542. MR1236836 https://doi.org/10.2307/2298123

NICOLA, M., ALSAFI, Z., SOHRABI, C., KERWAN, A., AL-JABIR, A., IOSIFIDIS, C., AGHA, M. and AGHA, R. (2020). The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *Int. J. Surg.* **78** 185–193. https://doi.org/10.1016/j.ijsu.2020.04.018

RIVERS, D. and VUONG, Q. (2002). Model selection tests for nonlinear dynamic models. *Econom. J.* **5** 1–39. MR1909299 https://doi.org/10.1111/1368-423X.t01-1-00071

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

VUONG, Q. H. (1989). Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica* **57** 307–333. MR0996939 https://doi.org/10.2307/1912557

WANG, W. and YU, J. (2015). Estimation of spatial panel data models with time varying spatial weights matrices. *Econom. Lett.* **128** 95–99. MR3315109 https://doi.org/10.1016/j.econlet.2015.01.021

WANG, X., PEI, T., LIU, Q., SONG, C., LIU, Y., CHEN, X., MA, J. and ZHANG, Z. (2020). Quantifying the time-lag effects of human mobility on the COVID-19 transmission: A multi-city study in China. *IEEE Access* **8** 216752–216761.

ZHANG, B., HENG, S., YE, T. and SMALL, D. S. (2023). Social distancing and COVID-19: Randomization inference for a structured dose-response relationship. *Ann. Appl. Stat.* **17** 23–46. MR4539020 https://doi.org/10.1214/22-aoas1613

ZHANG, D., FENG, L., WU, Y., LAN, W. and ZHOU, J. (2025). Supplement to "Temporal Network Influence Model with Application to the COVID-19 Population Flow Network." https://doi.org/10.1214/25-AOAS2025SUPP

ZHANG, Q. and WANG, H. (2011). On BIC's selection consistency for discriminant analysis. *Statist. Sinica* **21** 731–740. MR2829853 https://doi.org/10.5705/ss.2011.032a

ZHOU, J., TU, Y., CHEN, Y. and WANG, H. (2017). Estimating spatial autocorrelation with sampled network data. *J. Bus. Econom. Statist.* **35** 130–138. MR3591541 https://doi.org/10.1080/07350015.2015.1061437

ZHU, X., HUANG, D., PAN, R. and WANG, H. (2020). Multivariate spatial autoregressive model for large scale social networks. *J. Econometrics* **215** 591–606. MR4076946 https://doi.org/10.1016/j.jeconom.2018.11.018

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1403–1425
https://doi.org/10.1214/25-AOAS2030

THOMPSON SAMPLING FOR ZERO-INFLATED COUNT OUTCOMES WITH AN APPLICATION TO THE DRINK LESS MOBILE HEALTH STUDY

BY XUEQING LIU[1,a], NINA DELIU[2,c], TANUJIT CHAKRABORTY[3,d], LAUREN BELL[4,e] AND BIBHAS CHAKRABORTY[1,b]

[1]*Centre for Quantitative Medicine, Duke-NUS Medical School,* [a]*xueqing_liu@u.duke.nus.edu,* [b]*bibhas.chakraborty@duke-nus.edu.sg*

[2]*Department of Methods and Models for Economics, Territory and Finance, Sapienza University of Rome,* [c]*nina.deliu@uniroma1.it*

[3]*Department of Science and Engineering, Sorbonne University Abu Dhabi,* [d]*tanujit.chakraborty@sorbonne.ae*

[4]*Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London,* [e]*lauren.m.bell@kcl.ac.uk*

Mobile health (mHealth) interventions often aim to improve distal outcomes, such as clinical conditions, by optimizing proximal outcomes through just-in-time adaptive interventions. Contextual bandits provide a suitable framework for customizing such interventions according to individual time-varying contexts. However, unique challenges, such as modeling count outcomes within bandit frameworks, have hindered the widespread application of contextual bandits to mHealth studies. The current work addresses this challenge by leveraging count data models into online decision-making approaches. Specifically, we combine four common offline count data models (Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial regressions) with Thompson sampling, a popular contextual bandit algorithm. The proposed algorithms are motivated by and evaluated on a real dataset from the Drink Less trial, where they are shown to improve user engagement with the mHealth platform. The proposed methods are further evaluated on simulated data, achieving improvement in maximizing cumulative proximal outcomes over existing algorithms. Theoretical results on regret bounds are also derived. The `countts` R package provides an implementation of our approach.

REFERENCES

ABEILLE, M. and LAZARIC, A. (2017). Linear Thompson sampling revisited. In *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics* 176–184. PMLR.

AGRAWAL, S. and GOYAL, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning* 127–135. PMLR.

AGRAWAL, S. and JIA, R. (2017). Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds. In *Advances in Neural Information Processing Systems* 1184–1194.

AGUILERA, A., FIGUEROA, C. A., HERNANDEZ-RAMOS, R., SARKAR, U., CEMBALLI, A., GOMEZ-PATHAK, L., MIRAMONTES, J., YOM-TOV, E., CHAKRABORTY, B. et al. (2020). mHealth app using machine learning to increase physical activity in diabetes and depression: Clinical trial protocol for the DIAMANTE study. *BMJ Open* **10** e034723. https://doi.org/10.1136/bmjopen-2019-034723

BASTANI, H. and BAYATI, M. (2020). Online decision making with high-dimensional covariates. *Oper. Res.* **68** 276–294. MR4059503 https://doi.org/10.1287/opre.2019.1902

BEINEMA, T., OP DEN AKKER, H., HURMUZ, M., JANSEN-KOSTERINK, S. and HERMENS, H. (2022). Automatic topic selection for long-term interaction with embodied conversational agents in health coaching: A micro-randomized trial. *Internet Interv.* **27** 100502. https://doi.org/10.1016/j.invent.2022.100502

BELL, L., GARNETT, C., BAO, Y., CHENG, Z., QIAN, T., PERSKI, O., POTTS, H. W. W. and WILLIAMSON, E. (2023). How notifications affect engagement with a behavior change app: Results from a micro-randomized trial. *JMIR mHealth uHealth* **11** e38342. https://doi.org/10.2196/38342

Key words and phrases. Count data, contextual bandits, mobile health, online decision-making, Thompson sampling.

BELL, L., GARNETT, C., QIAN, T., PERSKI, O., POTTS, H. W., WILLIAMSON, E. et al. (2020). Notifications to improve engagement with an alcohol reduction app: Protocol for a micro-randomized trial. *JMIR Res. Protoc.* **9** e18690. https://doi.org/10.2196/18690

BOHN, M. J., BABOR, T. F. and KRANZLER, H. R. (1995). The alcohol use disorders identification test (AUDIT): Validation of a screening instrument for use in medical settings. *J. Stud. Alcohol* **56** 423–432. https://doi.org/10.15288/jsa.1995.56.423

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

CHAKRABORTY, S., ROY, S. and TEWARI, A. (2023). Thompson sampling for high-dimensional sparse linear contextual bandits. In *International Conference on Machine Learning* 3979–4008. PMLR.

CHAPELLE, O. and LI, L. (2011). An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems* 2249–2257.

DELIU, N., DAS, R., MAY, A., STEELE, J., DUCKWORTH, M., JONES, R. J., WILKINS, M. R., TOSHNER, M. and VILLAR, S. S. (2023). StratosPHere 2: A response-adaptive randomised placebo-controlled Phase II trial to evaluate hydroxychloroquine and phenylbutyrate in pulmonary arterial hypertension caused by mutations in BMPR2. *Research Square*. Preprint.

DELLA VECCHIA, R. and BASU, D. (2023). Stochastic online instrumental variable regression: Regrets for endogeneity and bandit feedback. arXiv e-prints arXiv–2302.

DING, Q., HSIEH, C.-J. and SHARPNACK, J. (2021). An efficient algorithm for generalized linear bandit: Online stochastic gradient descent and Thompson sampling. In *Proceedings of the* 24*th International Conference on Artificial Intelligence and Statistics* 1585–1593. PMLR.

DUMITRASCU, B., FENG, K. and ENGELHARDT, B. (2018). PG-TS: Improved Thompson sampling for logistic contextual bandits. In *Advances in Neural Information Processing Systems* 4624–4633.

FENG, C. X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *J. Stat. Distrib. Appl.* **8** 8. https://doi.org/10.1186/s40488-021-00121-4

FILIPPI, S., CAPPE, O., GARIVIER, A. and SZEPESVÁRI, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems* 586–594.

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

GARNETT, C., CRANE, D., WEST, R., BROWN, J. and MICHIE, S. (2019). The development of Drink Less: An alcohol reduction smartphone app for excessive drinkers. *Transl. Behav. Med.* **9** 296–307. https://doi.org/10.1093/tbm/iby043

GARNETT, C., PERSKI, O., MICHIE, S., WEST, R., FIELD, M., KANER, E., MUNAFÒ, M. R., GREAVES, F., HICKMAN, M. et al. (2021). Refining the content and design of an alcohol reduction app, Drink Less, to improve its usability and effectiveness: A mixed methods approach. F1000Research 10. https://doi.org/10.12688/f1000research.51416.2

GOURIÉROUX, C., MONFORT, A. and TROGNON, A. (1984). Pseudomaximum likelihood methods: Applications to Poisson models. *Econometrica* **52** 701–720. MR0740306 https://doi.org/10.2307/1913472

GREENEWALD, K., TEWARI, A., MURPHY, S. and KLASNJA, P. (2017). Action centered contextual bandits. In *Advances in Neural Information Processing Systems* 5979–5987.

HAMIDI, N. and BAYATI, M. (2020). On worst-case regret of linear Thompson sampling. arXiv preprint arXiv:2006.06790.

HUCH, E. K., SHI, J., ABBOTT, M. R., GOLBUS, J. R., MORENO, A. and DEMPSEY, W. H. (2023). Debiased machine learning and network cohesion for doubly-robust differential reward models in contextual bandits. arXiv preprint arXiv:2312.06403.

KLASNJA, P., HEKLER, E. B., SHIFFMAN, S., BORUVKA, A., ALMIRALL, D., TEWARI, A. and MURPHY, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology* **34** 1220. https://doi.org/10.1037/hea0000305

KUMAR, H., LI, T., SHI, J., MUSABIROV, I., KORNFIELD, R., MEYERHOFF, J., BHATTACHARJEE, A., KARR, C., NGUYEN, T. et al. (2024). Using adaptive bandit experiments to increase and investigate engagement in mental health. In *Proceedings of the AAAI Conference on Artificial Intelligence* **38** 22906–22912.

KVETON, B., ZAHEER, M., SZEPESVARI, C., LI, L., GHAVAMZADEH, M. and BOUTILIER, C. (2020). Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics* 2066–2076. PMLR.

LAMBERT, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** 1–14. https://doi.org/10.1080/00401706.1992.10485228

LI, L., CHU, W., LANGFORD, J. and SCHAPIRE, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the* 19*th International Conference on World Wide Web* 661–670. Association for Computing Machinery, New York, NY, USA.

LI, L., LU, Y. and ZHOU, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning* 2071–2080. PMLR.

LIAO, P., GREENEWALD, K., KLASNJA, P. and MURPHY, S. (2020). Personalized HeartSteps: A reinforcement learning algorithm for optimizing physical activity. *Proc*. *ACM Interact*. *Mob. Wearable Ubiquitous Technol*. **4** 1–22. https://doi.org/10.1145/3381007

LIU, X., DELIU, N. and CHAKRABORTY, B. (2023). Microrandomized trials: Developing just-in-time adaptive interventions for better public health. *Amer. J*. *Publ*. *Health* **113** 60–69. https://doi.org/10.2105/AJPH.2022.307150

LIU, X., DELIU, N., CHAKRABORTY, T., BELL, L. and CHAKRABORTY, B. (2025). Supplement to "Thompson sampling for zero-inflated count outcomes with an application to the Drink Less mobile health study." https://doi.org/10.1214/25-AOAS2030SUPPA, https://doi.org/10.1214/25-AOAS2030SUPPB, https://doi.org/10.1214/25-AOAS2030SUPPC

LIU, X., QIAN, T., BELL, L. and CHAKRABORTY, B. (2024). Incorporating nonparametric methods for estimating causal excursion effects in mobile health with zero-inflated count outcomes. *Biometrics* **80** Paper No. ujae054, 12. MR4755058 https://doi.org/10.1093/biomtc/ujae054

MCCULLAGH, P. and NELDER, J. A. (2019). *Generalized Linear Models*. Routledge. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

NAHUM-SHANI, I., SMITH, S. N., SPRING, B. J., COLLINS, L. M., WITKIEWITZ, K., TEWARI, A. and MURPHY, S. A. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Ann*. *Behav. Med*. **52** 446–462. https://doi.org/10.1007/s12160-016-9830-8

OH, M.-H. and IYENGAR, G. (2019). Thompson sampling for multinomial logit contextual bandits. In *Advances in Neural Information Processing Systems* 3151–3161.

PAYNE, E. H., HARDIN, J. W., EGEDE, L. E., RAMAKRISHNAN, V., SELASSIE, A. and GEBREGZIABHER, M. (2017). Approaches for dealing with various sources of overdispersion in modeling count data: Scale adjustment versus modeling. *Stat*. *Methods Med*. *Res*. **26** 1802–1823. MR3687179 https://doi.org/10.1177/0962280215588569

QIAN, T., WALTON, A. E., COLLINS, L. M., KLASNJA, P., LANZA, S. T., NAHUM-SHANI, I., RABBI, M., RUSSELL, M. A., WALTON, M. A. et al. (2022). The microrandomized trial for developing digital interventions: Experimental design and data analysis considerations. *Psychol*. *Methods*. https://doi.org/10.1037/met0000283

RUSSO, D. and VAN ROY, B. (2014). Learning to optimize via posterior sampling. *Math*. *Oper. Res*. **39** 1221–1243. MR3279764 https://doi.org/10.1287/moor.2014.0650

RUSSO, D. and VAN ROY, B. (2016). An information-theoretic analysis of Thompson sampling. *J*. *Mach*. *Learn*. *Res*. **17** Paper No. 68, 30. MR3517091

RUSSO, D. J., VAN ROY, B., KAZEROUNI, A., OSBAND, I., WEN, Z. et al. (2018). A tutorial on Thompson sampling. *Found*. *Trends Mach*. *Learn*. **11** 1–96. https://doi.org/10.1561/2200000070

SAUNDERS, J. B., AASLAND, O. G., BABOR, T. F., DE LA FUENTE, J. R. and GRANT, M. (1993). Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction* **88** 791–804. https://doi.org/10.1111/j.1360-0443.1993.tb02093.x

SCOTT, S. L. (2010). A modern Bayesian look at the multi-armed bandit. *Appl*. *Stoch*. *Models Bus*. *Ind*. **26** 639–658. MR2752378 https://doi.org/10.1002/asmb.874

SPRUIJT-METZ, D., MARLIN, B. M., PAVEL, M., RIVERA, D. E., HEKLER, E., DE LA TORRE, S., EL MISTIRI, M., GOLASZWESKI, N. M., LI, C. et al. (2022). Advancing behavioral intervention and theory development for mobile health: The HeartSteps II protocol. *Int*. *J*. *Environ*. *Res*. *Public Health* **19** 2267. https://doi.org/10.3390/ijerph19042267

SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning*: *An Introduction*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3889951

TEWARI, A. and MURPHY, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health* (J. M. Rehg, S. A. Murphy and S. Kumar, eds.) 495–517. Springer, Cham.

THOMPSON, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25** 285–294. https://doi.org/10.1093/biomet/25.3-4.285

TOMKINS, S., LIAO, P., KLASNJA, P. and MURPHY, S. (2021). IntelligentPooling: Practical Thompson sampling for mHealth. *Mach*. *Learn*. **110** 2685–2727. MR4306062 https://doi.org/10.1007/s10994-021-05995-8

TRELLA, A. L., ZHANG, K. W., NAHUM-SHANI, I., SHETTY, V., DOSHI-VELEZ, F. and MURPHY, S. A. (2022). Designing reinforcement learning algorithms for digital interventions: Pre-implementation guidelines. *Algorithms* **15** 255. https://doi.org/10.3390/a15080255

TRELLA, A. L., ZHANG, K. W., NAHUM-SHANI, I., SHETTY, V., DOSHI-VELEZ, F. and MURPHY, S. A. (2023). Reward design for an online reinforcement learning algorithm supporting oral self-care. In *Proceedings of the AAAI Conference on Artificial Intelligence* **37** 15724–15730. https://doi.org/10.1609/aaai.v37i13.26866

VAN ZWET, E. (2019). A default prior for regression coefficients. *Stat. Methods Med. Res.* **28** 3799–3807. MR4003623 https://doi.org/10.1177/0962280218817792

VER HOEF, J. M. and BOVENG, P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology* **88** 2766–2772. https://doi.org/10.1890/07-0043.1

XU, P., ZHENG, H., MAZUMDAR, E. V., AZIZZADENESHELI, K. and ANANDKUMAR, A. (2022). Langevin Monte Carlo for contextual bandits. In *Proceedings of the* 39*th International Conference on Machine Learning* 24830–24850. PMLR.

YAO, J., BRUNSKILL, E., PAN, W., MURPHY, S. and DOSHI-VELEZ, F. (2021). Power constrained bandits. In *Proceedings of the* 6*th Machine Learning for Healthcare Conference* **149** 209–259. PMLR.

ZHANG, J., CHEN, Y. and SINGH, A. (2022). Causal bandits: Online decision-making in endogenous settings. arXiv preprint arXiv:2211.08649.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1426–1447
https://doi.org/10.1214/24-AOAS1984

CENSORED C-LEARNING FOR DYNAMIC TREATMENT REGIME IN COLORECTAL CANCER STUDY

BY ZISHU ZHAN[1,a], ZHISHUAI LIU[2,b], CUNJIE LIN[3,c], DANHUI YI[4,d], JIAN LIU[5,e] AND YUFEI YANG[5,f]

[1]*Department of Biostatistics, School of Public Health, Southern Medical University,* [a]*zishu927@hotmail.com*
[2]*Department of Biostatistics & Bioinformatics, Duke University,* [b]*zhishuai.liu@duke.edu*
[3]*Center for Applied Statistics and School of Statistics, Renmin University of China,* [c]*lincunjie@ruc.edu.cn*
[4]*School of Statistics, Renmin University of China,* [d]*xueyi905@aliyun.com*
[5]*Xiyuan Clinical Medical College, Beijing University of Chinese Medicine,* [e]*xyyylj@126.com,* [f]*yyf93@vip.sina.com*

Dynamic treatment regimes (DTRs) represent sequential decision rules for multiple intervention stages. Each rule maps patients' covariates to optional treatments. The optimal dynamic treatment regime is the one that maximizes the mean outcome of interest if followed by the overall population. Motivated by a clinical study on the treatment of advanced colorectal cancer with traditional Chinese medicine, we propose a censored C-learning (CC-learning) method to estimate the DTR with multiple treatments based on survival data. To address the challenges of multiple stages with right censoring, we modified the backward recursion algorithm to adapt to the flexible number and timing of treatments. We propose a framework for multiple treatments that transforms the optimization problem of multiple treatment comparisons into an example-dependent, cost-sensitive classification problem. With data space expansion and classification techniques, the CC-learning method can produce an interpretable optimal DTR. We theoretically prove the method's optimality and assess its performance with finite sample simulations. Using our method, we identify the interpretable tree treatment regimes at each stage for the advanced colorectal cancer treatment data from Xiyuan Hospital.

REFERENCES

ABE, N., ZADROZNY, B. and LANGFORD, J. (2004). An iterative method for multi-class cost-sensitive learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* ACM, New York, NY, USA.

BELLMAN, R. (1966). Dynamic programming. *Science* **153** 34–37. MR0191704

BRULE, S. Y., JONKER, D. J., KARAPETIS, C. S., O'CALLAGHAN, C. J., MOORE, M. J., WONG, R., TEBBUTT, N. C., UNDERHILL, C., YIP, D. et al. (2015). Location of colon cancer (right-sided versus left-sided) as a prognostic factor and a predictor of benefit from cetuximab in NCIC CO.17. *Eur. J. Cancer* **51** 1405–1414.

CHEN, S. and ZHANG, B. (2023). Estimating and improving dynamic treatment regimes with a time-varying instrumental variable. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 427–453. MR4726972 https://doi.org/10.1093/jrsssb/qkad011

CHO, H., HOLLOWAY, S. T., COUPER, D. J. and KOSOROK, M. R. (2023). Multi-stage optimal dynamic treatment regimes for survival outcomes with dependent censoring. *Biometrika* **110** 395–410. MR4588360 https://doi.org/10.1093/biomet/asac047

CUI, Y. and TCHETGEN TCHETGEN, E. (2021). A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *J. Amer. Statist. Assoc.* **116** 162–173. MR4227683 https://doi.org/10.1080/01621459.2020.1783272

ELKAN, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the* 17*th International Joint Conference on Artificial Intelligence* Morgan Kaufmann, San Francisco, CA, USA.

GOLDBERG, Y. and KOSOROK, M. R. (2012). Q-learning with censored data. *Ann. Statist.* **40** 529–560. MR3014316 https://doi.org/10.1214/12-AOS968

Key words and phrases. Dynamic treatment regime, multiple treatments, classification perspective, survival data.

HUANG, X., CHOI, S., WANG, L. and THALL, P. F. (2015). Optimization of multi-stage dynamic treatment regimes utilizing accumulated data. *Stat. Med.* **34** 3423–3443. MR3412642 https://doi.org/10.1002/sim.6558

LI, Z., SHI, Q., LIU, M., JIA, L., HE, B., YANG, Y., LIU, J., LIN, H., LIN, H. K. et al. (2017). Validation and application of the MD Anderson symptom inventory for traditional Chinese medicine (MDASI-TCM). *J. Natl. Cancer Inst. Monograpr.* **2017** 48–57.

LIN, H. T. (2019). Advances in cost-sensitive multiclass and multilabel classification. In *Proceedings of the* 25*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* ACM, New York, NY, USA.

LUCKETT, D. J., LABER, E. B., KAHKOSKA, A. R., MAAHS, D. M., MAYER-DAVIS, E. and KOSOROK, M. R. (2020). Estimating dynamic treatment regimes in mobile health using V-learning. *J. Amer. Statist. Assoc.* **115** 692–706. MR4107673 https://doi.org/10.1080/01621459.2018.1537919

MA, H., ZENG, D. and LIU, Y. (2023). Learning optimal group-structured individualized treatment rules with many treatments. *J. Mach. Learn. Res.* **24** Paper No. [102], 48. MR4583263

OKEN, M. M., CREECH, R. H., TORMEY, D. C., HORTON, J., DAVIS, T. E., MCFADDEN, E. T. and CARBONE, P. P. (1982). Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Amer. J. Clin. Oncol.* **5** 649–655.

PRICE, T. J., BEEKE, C., ULLAH, S., PADBURY, R., MADDERN, G., RODER, D., TOWNSEND, A. R., MOORE, J., ROY, A. et al. (2015). Does the primary site of colorectal cancer impact outcomes for patients with metastatic disease? *Cancer* **121** 830–835. https://doi.org/10.1002/cncr.29129

SCHULTE, P. J., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2014). Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statist. Sci.* **29** 640–661. MR3300363 https://doi.org/10.1214/13-STS450

SHAO, C., ZUO, Q., LIN, J., YU, R. J., FU, Y., XIAO, M., SUN, L. L. and LIN, L. (2019). Effect of Chinese herbal medicine on the survival of colorectal cancer patients with liver-limited metastases: A retrospective cohort study, 2008 to 2017. *Integr. Cancer Ther.* **18** 1–10.

SIEGEL, R. L., GIAQUINTO, A. N. and JEMAL, A. (2024). Cancer statistics, 2024. *CA Cancer J. Clin.* **74** 12–49. https://doi.org/10.3322/caac.21820

SIMONEAU, G., MOODIE, E. E. M., NIJJAR, J. S., PLATT, R. W. and THE SCOTTISH EARLY RHEUMATOID ARTHRITIS INCEPTION COHORT INV (2020). Estimating optimal dynamic treatment regimes with survival outcomes. *J. Amer. Statist. Assoc.* **115** 1531–1539. MR4143483 https://doi.org/10.1080/01621459.2019.1629939

TAO, Y. and WANG, L. (2017). Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. *Biometrics* **73** 145–155. MR3632360 https://doi.org/10.1111/biom.12539

WALLACE, M. P. and MOODIE, E. E. M. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics* **71** 636–644. MR3402599 https://doi.org/10.1111/biom.12306

WANG, L., ZHOU, Y., SONG, R. and SHERWOOD, B. (2018). Quantile-optimal treatment regimes. *J. Amer. Statist. Assoc.* **113** 1243–1254. MR3862354 https://doi.org/10.1080/01621459.2017.1330204

WANG, Y., LIU, P., FANG, Y., TIAN, J., LI, S., XU, J., ZHAO, F., YIN, X., ZHANG, Q. et al. (2020). The effect of long-term traditional Chinese medicine treatment on survival time of colorectal cancer based on propensity score matching: A retrospective cohort study. *Evid.-Based Complement. Altern. Med.* **2020** 1–12.

XIA, J., ZHAN, Z. and ZHANG, J. (2024). Doubly robust estimation of optimal treatment regimes for survival data using an instrumental variable. *Stat. Comput.* **34** Paper No. 96, 18. MR4718734 https://doi.org/10.1007/s11222-024-10407-7

XUE, F., ZHANG, Y., ZHOU, W., FU, H. and QU, A. (2022). Multicategory angle-based learning for estimating optimal dynamic treatment regimes with censored data. *J. Amer. Statist. Assoc.* **117** 1438–1451. MR4480723 https://doi.org/10.1080/01621459.2020.1862671

YEH, M. H., CHIU, H. P., WU, M. C., KOO, M., LIN, N. W., LIAO, K. K., YEH, C. C. and LI, T. M. (2020). Integrated Chinese herbal medicine and western medicine on the survival in patients with colorectal cancer: A retrospective study of medical records. *Evid.-Based Complement. Altern. Med.* **2020** 1–10.

ZHAN, Z., LIU, Z., LIN, C., YI, D., LIU, J. and YANG, Y. (2025). Supplement to "Censored C-learning for Dynamic Treatment Regime in Colorectal Cancer Study." https://doi.org/10.1214/24-AOAS1984SUPPA, https://doi.org/10.1214/24-AOAS1984SUPPB

ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* **100** 681–694. MR3094445 https://doi.org/10.1093/biomet/ast014

ZHANG, B. and ZHANG, M. (2018). C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics* **74** 891–899. MR3860710 https://doi.org/10.1111/biom.12836

ZHANG, T., HE, W., ZI, M., SONG, G., YI, D. and YANG, Y. (2018). Cohort study on prognosis of patients with metastatic colorectal cancer treated with integrated Chinese and western medicine. *Chin. J. Integr. Med.* **24** 573–578.

ZHANG, Z., YI, D. and FAN, Y. (2022). Doubly robust estimation of optimal dynamic treatment regimes with multicategory treatments and survival outcomes. *Stat*. *Med*. **41** 4903–4923. MR4497087

ZHAO, Y.-Q., ZENG, D., LABER, E. B. and KOSOROK, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J*. *Amer. Statist*. *Assoc*. **110** 583–598. MR3367249 https://doi.org/10.1080/01621459.2014.937488

ZHOU, Y., WANG, L., SONG, R. and ZHAO, T. (2023). Transformation-invariant learning of optimal individualized decision rules with time-to-event outcomes. *J*. *Amer. Statist*. *Assoc*. **118** 2632–2644. MR4681609 https://doi.org/10.1080/01621459.2022.2068420

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1448–1472
https://doi.org/10.1214/24-AOAS1987

NONPARAMETRIC ADDITIVE VALUE FUNCTIONS: INTERPRETABLE REINFORCEMENT LEARNING WITH AN APPLICATION TO SURGICAL RECOVERY

BY PATRICK EMEDOM-NNAMDI[1,a], TIMOTHY R. SMITH[2,d], JUKKA-PEKKA ONNELA[1,b] AND JUNWEI LU[1,c]

[1]*Department of Biostatistics, Harvard University,* [a]*patrickemedom@hsph.harvard.edu,* [b]*onnela@hsph.harvard.edu,* [c]*junweilu@hsph.harvard.edu*

[2]*Computational Neurosciences Outcomes Center, Department of Neurosurgery, Brigham and Women's Hospital, Harvard Medical School,* [d]*trsmith@bwh.harvard.edu*

We propose a nonparametric additive model for estimating interpretable value functions in reinforcement learning, with an application in optimizing postoperative recovery through personalized, adaptive recommendations. While reinforcement learning has achieved significant success in various domains, recent methods often rely on black-box approaches, such as neural networks, which hinder the examination of individual feature contributions to a decision-making policy. Our novel method offers a flexible technique for estimating action-value functions without explicit parametric assumptions, overcoming the limitations of the linearity assumption of classical algorithms. By incorporating local kernel regression and basis expansion, we obtain a sparse, additive representation of the action-value function, enabling local approximation and retrieval of nonlinear, independent contributions of select state features and the interactions between joint feature pairs. We validate our approach through a simulation study and apply it to spine disease recovery, uncovering recommendations aligned with clinical knowledge. This method bridges the gap between flexible machine learning techniques and the interpretability required in healthcare applications, paving the way for more personalized interventions.

REFERENCES

AHN, I. and PARK, J. (2011). Drug scheduling of cancer chemotherapy based on natural actor-critic approach. *Biosystems* **106** 121–129. https://doi.org/10.1016/J.BIOSYSTEMS.2011.07.005

ANTOS, A., SZEPESVÁRI, C. and MUNOS, R. (2007). Fitted Q-iteration in continuous action-space MDPs. *Adv. Neural Inf. Process. Syst.* **20**.

BERTSEKAS, D. P. (2011). Approximate policy iteration: A survey and some new methods. *J. Control Theory Appl.* **9** 310–335. MR2833999 https://doi.org/10.1007/s11768-011-1005-3

BOARO, A., REEDER, H. T. and SIDDI, F. (2021). Smartphone GPS signatures of patients undergoing spine surgery correlate with mobility and current gold standard outcome measures Digital Phenotyping in Neurosurgery View project Intracranial aneurysm View project. *J. Neurosurg., Spine* **35** 796–806. https://doi.org/10.3171/2021.2.SPINE202181

BOTHE, M. K., DICKENS, L., REICHEL, K., TELLMANN, A., ELLGER, B., WESTPHAL, M. and FAISAL, A. A. (2013). The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Expert Rev. Med. Devices* **10** 661–673. https://doi.org/10.1586/17434440.2013.827515

BOWYER, A. J. and ROYSE, C. F. (2016). Postoperative recovery and outcomes—What are we measuring and for whom? *Anaesthesia* **71** 72–77. https://doi.org/10.1111/anae.13312

BURGESS, L. C. and WAINWRIGHT, T. W. (2019). What is the evidence for early mobilisation in elective spine surgery? A narrative review. *Healthcare* **7** 92. https://doi.org/10.3390/healthcare7030092

CLARK, D. J. (2015). Automaticity of walking: Functional significance, mechanisms, measurement and rehabilitation strategies. *Front. Human Neurosci.* **9** 246. https://doi.org/10.3389/fnhum.2015.00246

COTE, D. J., BARNETT, I., ONNELA, J.-P. and SMITH, T. R. (2019). Digital phenotyping in patients with spine disease: A novel approach to quantifying mobility and quality of life. *World Neurosurg.* **126** e241–e249. https://doi.org/10.1016/j.wneu.2019.01.297

Key words and phrases. Reinforcement learning, interpretable value functions, nonparametric, digital health.

DUC, C., SALVIA, P., LUBANSU, A., FEIPEL, V. and AMINIAN, K. (2013). Objective evaluation of cervical spine mobility after surgery during free-living activity. *Clin. Biomech.* **28** 364–369. https://doi.org/10.1016/j.clinbiomech.2013.03.006

ELSARRAG, M., SOLDOZY, S., PATEL, P., NORAT, P., SOKOLOWSKI, J. D., PARK, M. S., TVRDIK, P. and KALANI, M. Y. S. (2019). Enhanced recovery after spine surgery: A systematic review. *Neurosurg. Focus* **46** 1–8. https://doi.org/10.3171/2019.1.FOCUS18700

EMEDOM-NNAMDI, P., SMITH, T. R., ONNELA, J.-P. and LU, J. (2025). Supplement to "Nonparametric additive value functions: Interpretable reinforcement learning with an application to surgical recovery." https://doi.org/10.1214/24-AOAS1987SUPPA, https://doi.org/10.1214/24-AOAS1987SUPPB

FAN, J. and JIANG, J. (2005). Nonparametric inferences for additive models. *J. Amer. Statist. Assoc.* **100** 890–907. https://doi.org/10.1198/016214504000001439

FARAHMAND, A., GHAVAMZADEH, M., SZEPESVÁRI, C. and MANNOR, S. (2016). Regularized policy iteration with nonparametric function spaces. *J. Mach. Learn. Res.* **17** Paper No. 139, 66 pp. MR3555030

GEIST, M. and SCHERRER, B. (2011). l1-penalized projected Bellman residual. Technical Report No. 11.

GHAVAMZADEH, M., MANNOR, S., PINEAU, J. and TAMAR, A. (2016). Bayesian reinforcement learning: A survey. *Found. Trends Mach. Learn.* **8** 359–483. https://doi.org/10.1561/2200000049

GOTTESMAN, O., JOHANSSON, F., KOMOROWSKI, M., FAISAL, A., SONTAG, D., DOSHI-VELEZ, F. and CELI, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nat. Med.* **25** 16–18. https://doi.org/10.1038/s41591-018-0310-5

HASTIE, T. J. (2017). Generalized additive models. In *Statistical Models in S* 249–307. https://doi.org/10.1201/9780203738535-7

HOFFMAN, M. W., LAZARIC, A., GHAVAMZADEH, M. and MUNOS, R. (2012). Regularized least squares temporal difference learning with nested ℓ_2 and ℓ_1 penalization. In *Recent Advances in Reinforcement Learning. EWRL* 2011. *Lecture Notes in Computer Science* **7188**. Springer, Berlin. https://doi.org/10.1007/978-3-642-29946-9_13

HOWARD, M. and NAKAMURA, Y. (2013). Locally weighted least squares policy iteration for model-free learning in uncertain environments. In *IEEE International Conference on Intelligent Robots and Systems* 1223–1229. https://doi.org/10.1109/IROS.2013.6696506

HOWARD, R. A. (1960). *Dynamic Programming and Markov Processes*. Wiley, New York. MR0118514

JAENSSON, M., DAHLBERG, K. and NILSSON, U. (2019). Factors influencing day surgery patients' quality of postoperative recovery and satisfaction with recovery: A narrative review. *Perioper. Med.* **8** 3. https://doi.org/10.1186/s13741-019-0115-1

KARAS, M., MARINSEK, N., GOLDHAHN, J., FOSCHINI, L., RAMIREZ, E. and CLAY, I. (2020). Predicting subjective recovery from lower limb surgery using consumer wearables. *Digit. Biomark.* **4** 73–86. https://doi.org/10.1159/000511531

KOLTER, J. Z. and NG, A. Y. (2009). Regularization and feature selection in least-squares temporal difference learning. In *ACM International Conference Proceeding Series* **382**. https://doi.org/10.1145/1553374.1553442

KUMAR, A., ZHOU, A., TUCKER, G. and LEVINE, S. (2020). Conservative Q-learning for offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2020-December**. https://doi.org/10.48550/arxiv.2006.04779

LAFFERTY, J. and WASSERMAN, L. (2008). Rodeo: Sparse, greedy nonparametric regression. *Ann. Statist.* **36** 28–63. MR2387963 https://doi.org/10.1214/009053607000000811

LAGOUDAKIS, M. G. and PARR, R. (2004). Least-squares policy iteration. *J. Mach. Learn. Res.* **4** 1107–1149. MR2125347 https://doi.org/10.1162/1532443041827907

LIU, G. and ONNELA, J.-P. (2021). Bidirectional imputation of spatial GPS trajectories with missingness using sparse online Gaussian process. *J. Amer. Med. Inform. Assoc.* **28** 1777–1784. https://doi.org/10.1093/jamia/ocab069

LU, J., KOLAR, M. and LIU, H. (2020). Kernel meets sieve: Post-regularization confidence bands for sparse additive model. *J. Amer. Statist. Assoc.* **115** 2084–2099. MR4189778 https://doi.org/10.1080/01621459.2019.1689984

MNIH, V., KAVUKCUOGLU, K., SILVER, D., GRAVES, A., ANTONOGLOU, I., WIERSTRA, D. and RIEDMILLER, M. (2013). Playing Atari with deep reinforcement learning. Preprint. Available at arXiv:1312.5602.

MNIH, V., KAVUKCUOGLU, K., SILVER, D., RUSU, A. A., VENESS, J., BELLEMARE, M. G., GRAVES, A., RIEDMILLER, M., FIDJELAND, A. K. et al. (2015). Human-level control through deep reinforcement learning. *Nature* **518** 529–533. https://doi.org/10.1038/nature14236

O'DONOGHUE, B., OSBAND, I., MUNOS, R. and MNIH, V. (2017). The uncertainty Bellman equation and exploration. In 35*th International Conference on Machine Learning*, *ICML* 2018 **9** 6154–6173.

ONNELA, J.-P. (2021). Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology* **46** 45–54. https://doi.org/10.1038/s41386-020-0771-3

OZKARA, G. O., OZGEN, M., OZKARA, E., ARMAGAN, O., ARSLANTAS, A. and ATASOY, M. A. (2015). Effectiveness of physical therapy and rehabilitation programs starting immediately after lumbar disc surgery. *Turk. Neurosurg.* **25** 372–379. https://doi.org/10.5137/1019-5149.JTN.8440-13.0

PADMANABHAN, R., MESKIN, N. and HADDAD, W. M. (2015). Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. *Biomed. Signal Process. Control* **22** 54–64. https://doi.org/10.1016/J.BSPC.2015.05.013

PANDA, N., SOLSKY, I., HAWRUSIK, B., LIU, G., REEDER, H., LIPSITZ, S., DESAI, E. V., LOWERY, K. W., MILLER, K. et al. (2020a). Smartphone global positioning system (GPS) data enhances recovery assessment after breast cancer surgery. *Ann. Surg. Oncol.* **28** 985–994. https://doi.org/10.1245/s10434-020-09004-5

PANDA, N., SOLSKY, I., HUANG, E. J., LIPSITZ, S., PRADARELLI, J. C., DELISLE, M., CUSACK, J. C., GADD, M. A., LUBITZ, C. C. et al. (2020b). Using smartphones to capture novel recovery metrics after cancer surgery. *JAMA Surg.* **155** 123–129. https://doi.org/10.1001/jamasurg.2019.4702

PENG, X., DING, Y., WIHL, D., GOTTESMAN, O., KOMOROWSKI, M., LEHMAN, L.-W. H., ROSS, A., FAISAL, A. and DOSHI-VELEZ, F. (2018). Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. *AMIA Annual Symp. Proc.* **2018** 887–896.

PEYSER, T., DASSAU, E., BRETON, M. and SKYLER, J. S. (2014). The artificial pancreas: Current status and future prospects in the management of diabetes. *Ann. N.Y. Acad. Sci.* **1311** 102–123. https://doi.org/10.1111/nyas.12431

RAGHU, A., KOMOROWSKI, M., CELI, L. A., SZOLOVITS, P. and GHASSEMI, M. (2017). Continuous state-space models for optimal sepsis treatment—A deep reinforcement learning approach 1–17.

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

RICHTÁRIK, P. and TAKÁČ, M. (2014). Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Math. Program.* **144** 1–38. MR3179953 https://doi.org/10.1007/s10107-012-0614-z

RIEDMILLER, M. (2005). Neural fitted Q iteration—First experiences with a data efficient neural reinforcement learning method. In *Machine Learning: ECML* 2005. *Lecture Notes in Computer Science* **3720**. Springer, Berlin. https://doi.org/10.1007/11564096_32

SINZINGER, E. D. and MOORE, B. (2011). Sedation of simulated Icu patients using reinforcement learning based control. *Int. J. Artif. Intell. Tools* **14** 137–156. https://doi.org/10.1142/S021821300500203X

SONABEND-W, A., LU, J., CELI, L. A., CAI, T. and SZOLOVITS, P. (2020). Expert-supervised reinforcement learning for offline policy learning and evaluation. *Adv. Neural Inf. Process. Syst.* **2020-December**.

SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning: An Introduction*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3889951

TOROUS, J., STAPLES, P. and ONNELA, J.-P. (2015). Realizing the potential of mobile mental health: New methods for new data in psychiatry. *Curr. Psychiatry Rep.* **17** 602. https://doi.org/10.1007/s11920-015-0602-0

TZIORTZIOTIS, N. and DIMITRAKAKIS, C. (2017). Bayesian inference for least squares temporal difference regularization. Technical Report.

VAN HASSELT, H. (2010). Double Q-learning. *Adv. Neural Inf. Process. Syst.* **23**.

WAINWRIGHT, T. W., IMMINS, T. and MIDDLETON, R. G. (2016). Enhanced recovery after surgery (ERAS) and its applicability for major spine surgery. *Bailliere's Best Pract. Res., Clin. Anaesthesiol.* **30** 91–102. https://doi.org/10.1016/j.bpa.2015.11.001

WANG, Z., NOVIKOV, A., ZOLNA, K., SPRINGENBERG, J. T., REED, S., SHAHRIARI, B., SIEGEL, N., MEREL, J., GULCEHRE, C. et al. (2020). Critic regularized regression. *Adv. Neural Inf. Process. Syst.* **2020-December**.

XU, X., HU, D. and LU, X. (2007). Kernel-based least squares policy iteration for reinforcement learning. *IEEE Trans. Neural Netw.* **18** 973–992. https://doi.org/10.1109/TNN.2007.899161

YEZIERSKI, R. P. (2012). The effects of age on pain sensitivity: Preclinical studies. *Pain Med.* **13 Suppl 2** 27–36. https://doi.org/10.1111/j.1526-4637.2011.01311.x

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1473–1493
https://doi.org/10.1214/24-AOAS1962

A BAYESIAN APPROACH TO GRAPPA PARALLEL FMRI IMAGE RECONSTRUCTION INCREASES SNR AND POWER OF TASK DETECTION

BY CHASE J. SAKITIS[a] AND DANIEL B. ROWE[b]

Mathematical and Statistical Sciences, Marquette University, [a]*chase.sakitis@marquette.edu,* [b]*daniel.rowe@marquette.edu*

In fMRI, capturing brain activation during a task is dependent on how quickly k-space arrays are obtained. Acquiring full k-space arrays, which are reconstructed into images using the inverse Fourier transform (IFT), that make up volume images can take a considerable amount of scan time. Undersampling k-space reduces the acquisition time but results in aliased, or "folded," images. GeneRalized Autocalibrating Partial Parallel Acquisition (GRAPPA) is a parallel imaging technique that yields full images from subsampled arrays of k-space. GRAPPA uses localized interpolation weights, which are estimated prescan and fixed over time, to fill in the missing spatial frequencies of the subsampled k-space. Here we propose a Bayesian approach to GRAPPA (BGRAPPA) where prior distributions for the unacquired spatial frequencies, localized interpolation weights, and k-space measurement uncertainty are assessed from the a priori calibration k-space arrays. The prior information is utilized to estimate the missing spatial frequency values from the posterior distribution and reconstruct into full field-of-view images. Our BGRAPPA technique successfully reconstructed both a simulated and experimental time series resulting in reduced noise leading to an increased signal-to-noise ratio (SNR) and stronger power of task detection.

REFERENCES

ATKINSON, D., HILL, D. L. G., STOYLE, P. N. R., SUMMERS, P. E. and KEEVIL, S. F. (1997). Automatic correction of motion artifacts in magentic resonance images using an entropy focus criterion. *IEEE Trans. Med. Imag.* **16** 903–910.

BANDETTINI, P., JESMANOWICZ, A., WONG, E. and HYDE, J. S. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magn. Reson. Med.* **30** 161–173.

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

GENOVESE, C. R., LAZAR, N. A. and NICHOLS, T. E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* **15** 870–878.

GRISWOLD, M. A., JAMOB, P. M., HEIDEMANN, R. M., NITTKA, M., JELLUS, V., WANG, J., KIEFER, B. and HAASE, A. (2002). Generalized autocalibrating artially parallel acquisition (GRAPPA). *Magn. Reson. Med.* **47** 1202–1210.

HYDE, J. S., JESMANOWICZ, A., FRONCISZ, W., KNEELAND, J. B., GRIST, T. M. and CAMPAGNA, N. F. (1986). Parallel image acquisition from noninteracting local coils. *J. Magn. Reson.* (1969) **70** 512–517.

KARAMAN, M. M., BRUCE, I. P. and ROWE, D. B. (2014). A statistical fMRI model for differential T_2* contrast incorporating T_1 and T_2* of gray matter. *Magn. Reson. Imaging* **32** 9–27.

KARAMAN, M. M., BRUCE, I. P. and ROWE, D. B. (2015). Incorporating relaxivities to more accurately reconstruct MR images. *Magn. Reson. Imaging* **33** 374–384.

KORNAK, J., YOUNG, K., SCHUFF, N., DU, A., MAUDSLEY, A. A. and WEINER, M. W. (2010). K-Bayes reconstruction for perfusion MRI. I: Concepts and application. *J. Digit. Imag.* **23** 277–286.

KUMAR, A., WELTI, D. and ERNST, R. R. (1975). NMR Fourier zeugmatography. *J. Magn. Reson.* (1969) **18** 69–83.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34** 1–41. MR0415861

LINDQUIST, M. A. (2008). The statistical analysis of fMRI data. *Statist. Sci.* **23** 439–464. MR2530545 https://doi.org/10.1214/09-STS282

Key words and phrases. Bayesian, GRAPPA, fMRI, reconstruction.

LINDQUIST, M. A., ZHANG, C., GLOVER, G. and SHEPP, L. (2008). Rapid three-dimensional functional magnetic resonance imaging of the initial negative BOLD response. *J. Magn. Reson.* **191** 100–111.

LOGAN, B. R. and ROWE, D. B. (2004). An evaluation of thresholding techniques in fMRI analysis. *NeuroImage* **22** 95–108. https://doi.org/10.1016/j.neuroimage.2003.12.047

O'HAGAN, A. (1994). *Kendall's Advanced Theory of Statistics. Vol.* 2*B. Kendall's Library of Statistics*. Wiley, New York. MR1285356

OGAWA, S., LEE, T. M., NAYAK, A. S. and GLYNN, P. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn. Reson. Med.* **14** 68–78.

PRUESSMANN, K. P., WEIGER, M., SCHEIDEGGER, M. B. and BOESIGER, P. (1999). SENSE: Sensitivity encoding for fast MRI. *Magn. Reson. Med.* **42** 952–962.

ROWE, D. B. (2005). Modeling both the magnitude and phase of complex-valued fMRI data. *NeuroImage* **25** 1310–1324. https://doi.org/10.1016/j.neuroimage.2005.01.034

ROWE, D. B. (2009). Magnitude and phase signal detection in complex-valued fMRI data. *Magn. Reson. Med.* **62** 1356–1357.

ROWE, D. B. and LOGAN, B. R. (2004). A complex way to compute fMRI activation. *NeuroImage* **23** 1078–1092. https://doi.org/10.1016/j.neuroimage.2004.06.042

ROWE, D. B., MELLER, C. P. and HOFFMANN, R. G. (2007). Characterizing phase-only fMRI data with an angular regression model. *J. Neurosci. Methods* **161** 331–341. https://doi.org/10.1016/j.jneumeth.2006.10.024

SAKITIS, C. J., BROWN, D. A. and ROWE, D. B. (2025). A Bayesian complex-valued latent variable model applied to functional magnetic resonance imaging. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **74** 100–125. MR4863100 https://doi.org/10.1093/jrsssc/qlae046

SAKITIS, C. J. and ROWE, D. B. (2025). Supplement to "A Bayesian approach to GRAPPA parallel fMRI image reconstruction increases SNR and power of task detection." https://doi.org/10.1214/24-AOAS1962SUPP

STEINHOFF, S., ZAITSEV, M., ZILLES, K. and SHAH, N. J. (2001). Fast T_1 mapping with volume coverage. *Magn. Reson. Med.* **46** 131–140.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1494–1513
https://doi.org/10.1214/24-AOAS1981

RICE-DISTRIBUTED AUTOREGRESSIVE TIME SERIES MODELING OF MAGNITUDE FUNCTIONAL MRI DATA

BY DANIEL W. ADRIAN[1,a], RANJAN MAITRA[2,b] AND DANIEL B. ROWE[3,c]

[1]*Department of Statistics, Grand Valley State University,* [a]*adriand1@gvsu.edu*
[2]*Department of Statistics, Iowa State University,* [b]*maitra@iastate.edu*
[3]*Department of Mathematical and Statistical Sciences, Marquette University,* [c]*daniel.rowe@marquette.edu*

Functional magnetic resonance imaging (fMRI) data generally consist of time series image volumes of the magnitude of complex-valued observations at each voxel. However, incorporating Gaussian-based time series models and the Rice distribution—a more accurate model for the data—in the time series have been separated by a distributional "mismatch." We bridge this gap by including pth-order autoregressive (AR) errors into the Gaussian model for the latent real and imaginary components underlying the Rice-distributed magnitude data. Parameter estimation is then done by augmenting the observed magnitude data with the missing phase data in an expectation-maximization (EM) algorithm framework and followed by AR order determination and computation of test statistics for activation detection. Using simulated and experimental low-SNR fMRI data, we compare the performance of this Ricean time series model with a Gaussian AR(p) model for the magnitude data and also with a complex Gaussian time series model for the entire complex-valued data. Our results show improved parameter estimation and activation detection under the Ricean AR(p) model for the magnitude data than its Gaussian counterpart. The model using the complex-valued data (which is rarely collected in practice) detects activation better than both magnitude-only models but only because it has more data. Thus, while our results here provide for the improved analysis of commonly-collected and archived magnitude-only fMRI datasets, they also argue strongly against the currently routine practice of discarding the phase of the complex-valued fMRI time series, advocating instead for their inclusion in the analysis.

REFERENCES

ABDI, A., TEPEDELENLIOGLU, C., KAVEH, M. and GIANNAKIS, G. (2001). On the estimation of the K parameter for the Rice fading distribution. *IEEE Commun. Lett.* **5** 92–94.

ABRAMOWITZ, M. and STEGUN, I. A. (1966). *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*. Dover, New York. MR0208797

ADRIAN, D. W., MAITRA, R. and ROWE, D. B. (2013). Ricean over Gaussian modelling in magnitude fMRI analysis—added complexity with negligible practical benefits. *Stat* **2** 303–316. MR4027320 https://doi.org/10.1002/sta4.34

ADRIAN, D. W., MAITRA, R. and ROWE, D. B. (2018). Complex-valued time series modeling for improved activation detection in fMRI studies. *Ann. Appl. Stat.* **12** 1451–1478. MR3852684 https://doi.org/10.1214/17-AOAS1117

ADRIAN, D. W., MAITRA, R. and ROWE, D. B. (2025). Supplement to "Rice-distributed autoregressive time series modeling of magnitude functional MRI data." https://doi.org/10.1214/24-AOAS1981SUPPA, https://doi.org/10.1214/24-AOAS1981SUPPB

AITKIN, M. and AITKIN, I. (1996). A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. *Stat. Comput.* **6** 127–130.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) 267–281. Akad. Kiadó, Budapest. MR0483125

Key words and phrases. Bilateral finger-tapping motor experiment, EM algorithm, empirical information matrix, hemodynamic response function, Monte Carlo integration, Rice distribution, signal-to-noise ratio.

ARAFA, A. and MESSIER, G. G. (2010). A Gaussian model for dead-reckoning mobile sensor position error. In 2010 *IEEE* 72*nd Vehicular Technology Conference—Fall* 1–5.

BAGGIO, R. and MUZY, J.-F. (2024). Improving probabilistic wind speed forecasting using M-Rice distribution and spatial data integration. *Appl*. *Energy* **360** 122840.

BAÏLE, R., MUZY, J. F. and POGGI, P. (2011). An M-Rice wind speed frequency distribution. *Wind Energy* **14** 735–748.

BALCHANDANI, P. and NAIDICH, T. P. (2015). Ultra-high-field MR neuroimaging. *Am*. *J*. *Neuroradiol*. **36** 1204–1215. https://doi.org/10.3174/ajnr.A4180

BANDETTINI, P. A., JESMANOWICZ, A., WONG, E. C. and HYDE, J. S. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magn*. *Reson*. *Med*. **30** 161–173. https://doi.org/10.1002/mrm.1910300204

BAR-SHALOM, Y., LI, X. R. and KIRUBARAJAN, T. (2001). *Estimation with Applications to Tracking and Navigation*: *Theory Algorithms and Software*. Wiley, New York.

BELLIVEAU, J. W., KENNEDY, JR., D. N., MCKINSTRY, R. C., BUCHBINDER, B. R., WEISSKOFF, R. M., COHEN, M. S., VEVEA, J. M., BRADY, T. J. and ROSEN, B. R. (1991). Functional mapping of the human visual cortex by magnetic resonance imaging. *Science* **254** 716–719. https://doi.org/10.1126/science.1948051

BEST, D. J., RAYNER, J. C. W. and THAS, O. (2010). Easily applied tests of fit for the Rayleigh distribution. *Sankhya B* **72** 254–263. MR2820134 https://doi.org/10.1007/s13571-011-0011-2

BROWN, T. R., KINCAID, B. M. and UGURBIL, K. (1982). NMR chemical shift imaging in three dimensions. In *Proceedings of the National Academy of Sciences*, *USA* **79** 3523–3526.

BULLMORE, E., BRAMMER, M., WILLIAMS, S. C. R., RABE-HESKETH, S., JANOT, N., DAVID, A., MELLERS, J., HOWARD, R. and SHAM, P. (1996). Statistical methods of estimation and inference for function MR image analysis. *Magn*. *Reson*. *Med*. **35** 261–277.

CARPENTER, M. B. (1991). *Core Text of Neuroanatomy*. Williams & Wilkins, Baltimore.

CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference*, 2nd ed. Thomson Learning.

COSOTTINI, M. and ROCCATAGLIATA, L. (2021). Neuroimaging at 7 T: Are we ready for clinical transition? *Eur*. *Radiol*. *Exp*. **5** 37. https://doi.org/10.1186/s41747-021-00234-0

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

DEN DEKKER, A. J., POOT, D. H. J., BOS, R. and SIJBERS, J. (2009). Likelihood-based hypothesis tests for brain activation detection from MRI data disturbed by colored noise: A simulation study. *IEEE Trans*. *Med*. *Imag*. **28** 287–296. https://doi.org/10.1109/TMI.2008.2004427

EFRON, B. and TIBSHIRANI, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist*. *Sci*. **1** 54–77. MR0833275

FRISTON, K. J., FRITH, C. D., LIDDLE, P. F., DOLAN, R. J., LAMMERTSMA, A. A. and FRACKOWIAK, R. S. J. (1990). The relationship between global and local changes in PET scans. *J. Cereb. Blood Flow Metab.* **10** 458–466.

FRISTON, K. J., HOLMES, A. P., WORSLEY, K. J., POLINE, J. B., FRITH, C. D. and FRACKOWIAK, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Hum*. *Brain Mapp*. **2** 189–210.

FRISTON, K. J., JEZZARD, P. and TURNER, R. (1994). Analysis of functional MRI time-series. *Hum*. *Brain Mapp*. **1** 153–171.

FRISTON, K. J., JOSEPHS, O., ZARAHN, E., HOLMES, A. P., ROUQUETTE, S. and POLINE, J. (2000). To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis. *NeuroImage* **12** 196–208. https://doi.org/10.1006/nimg.2000.0609

GENOVESE, C. R., LAZAR, N. A. and NICHOLS, T. E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* **15** 870–878.

GLISSON, T. H. (2011). *Introduction to Circuit Analysis and Design*. Springer, The Netherlands.

GLOVER, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage* **9** 416–429. https://doi.org/10.1006/nimg.1998.0419

GUDBJARTSSON, H. and PATZ, S. (1995). The Rician distribution of noisy MRI data. *Magn*. *Reson*. *Med*. **34** 910–914. https://doi.org/10.1002/mrm.1910340618

HAHN, A. D., NENCKA, A. S. and ROWE, D. B. (2009). Improving robustness and reliability of phase-sensitive fMRI analysis using temporal off-resonance alignment of single-echo timeseries (TOAST). *NeuroImage* **44** 742–752. https://doi.org/10.1016/j.neuroimage.2008.10.001

HAJRI, N., YOUSSEF, N. and PATZOLD, M. (2009). A study on the statistical properties of double Hoyt fading channels. In 2009 6*th International Symposium on Wireless Communication Systems* 201–205.

HENKELMAN, R. M. (1985). Measurement of signal intensities in the presence of noise in MR images. *Med*. *Phys*. **12** 232–233. https://doi.org/10.1118/1.595711

Islam, K. T., Zhong, S., Zakavi, P., Chen, Z., Kavnoudias, H., Farquharson, S., Durbridge, G., Barth, M., McMahon, K. L. et al. (2023). Improving portable low-field MRI image quality through image-to-image translation using paired low- and high-field images. *Sci. Rep.* **13**.

Jain, A. K. (1989). *Fundaments of Digital Image Processing*. Prentice Hall, New York.

Jesmanowicz, A., Wong, E. C. and Hyde, J. S. (1993). Phase correction for EPI using internal reference lines. In *Proceedings from the. International Society of Magnetic Resonance in Medicine* **12** 1239.

Jezzard, P. and Clare, S. (2001). Principles of nuclear magnetic resonance and MRI. In *Functional MRI: An Introduction to Methods* (P. Jezzard, P. M. Matthews and S. M. Smith, eds.) **3** 67–92. Oxford Univ. Press, London.

Kumar, A., Welti, D. and Ernst, R. R. (1975). NMR Fourier Zeugmatography. *J. Magn. Reson.* **18** 69–83.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S. et al. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl. Acad. Sci. USA* **89** 5675–5679.

Lazar, N. A. (2008). *The Statistical Analysis of Functional MRI Data*. Springer, Berlin.

Lee, C. C., Jack, C. R. and Riederer, S. J. (1998). Mapping of the central sulcus with functional MR: Active versus passive activation tasks. *Am. J. Neuroradiol.* **19** 847–852.

Lee, C. C., Ward, H. A., Sharbrough, F. W., Meyer, F. B., Marsh, W. R., Raffel, C., So, E. L., Cascino, G. D., Shin, C. et al. (1999). Assessment of functional MR imaging in neurosurgical planning. *Am. J. Neuroradiol.* **20** 1511–1519.

Hahn, A. D., Nencka, A. S. and Rowe, D. B. (2012). Enhancing the utility of complex-valued functional magnetic resonance imaging detection of neurobiological processes through postacquisition estimation and correction of dynamic B_0 errors and motion. *Hum. Brain Mapp.* **33** 288–306.

Lindsey, W. (1964). Error probabilities for Rician fading multichannel reception of binary and n-ary signals. *IEEE Trans. Inf. Theory* **10** 339–350.

Liu, Y., Leong, A. T. L., Zhao, Y., Xiao, L., Mak, H. K. F., Tsang, A. C. O., Lau, G. K. K., Leung, G. K. K. and Wu, E. X. (2021). A low-cost and shielding-free ultra-low-field brain MRI scanner. *Nat. Commun.* **12** 1–14.

Ljunggren, S. (1983). A simple graphical representation of Fourier-based imaging methods. *J. Magn. Reson.* **54** 338–343.

Locascio, J. J., Jennings, P. J., Moore, C. I. and Corkin, S. (1997). Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. *Hum. Brain Mapp.* **5** 168–193. https://doi.org/10.1002/(SICI)1097-0193(1997)5:3<168::AID-HBM3>3.0.CO;2-1

Logan, B. R. and Rowe, D. B. (2004). An evaluation of thresholding techniques in fMRI analysis. *NeuroImage* **22** 95–108. https://doi.org/10.1016/j.neuroimage.2003.12.047

Macovski, A. (1996). Noise in MRI. *Magn. Reson. Med.* **36** 494–497. https://doi.org/10.1002/mrm.1910360327

Marchini, J. L. and Ripley, B. D. (2000). A new statistical approach to detecting significant activation in functional MRI. *NeuroImage* **12** 366–380. https://doi.org/10.1006/nimg.2000.0628

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Med. Decis. Mak.* **9** 190–195. https://doi.org/10.1177/0272989X8900900307

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR2392878 https://doi.org/10.1002/9780470191613

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *J. Roy. Statist. Soc. Ser. B* **51** 127–138. MR0984999

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

Miller, J. W. (1995). Exact maximum likelihood estimation in autoregressive processes. *J. Time Series Anal.* **16** 607–615. MR1366802 https://doi.org/10.1111/j.1467-9892.1995.tb00258.x

Murphy, K., Bodurka, J. and Bandettini, P. A. (2007). How long to scan? The relationship between fMRI temporal signal to noise ratio and necessary scan duration. *NeuroImage* **34** 565–574.

Nencka, A. S., Hahn, A. D. and Rowe, D. B. (2008). The use of three navigator echoes in Cartesian EPI reconstruction reduces Nyquist ghosting. In *Proceedings from the. International Society of Magnetic Resonance in Medicine* **16** 3032.

Ogawa, S., Lee, T. M., Nayak, A. S. and Glynn, P. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn. Reson. Med.* **14** 68–78.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol. I: Theory of Statistics* 697–715. Univ. California Press, Berkeley, CA. MR0400516

Pourahmadi, M. (2001). *Foundations of Time Series Analysis and Prediction Theory*. *Wiley Series in Probability and Statistics: Applied Probability and Statistics*. Wiley Interscience, New York. MR1849562

POWERS, J. M., IOACHIM, G. and STROMAN, P. W. (2018). Ten key insights into the use of spinal cord fMRI. *Brain Sci*. **8**. https://doi.org/10.3390/brainsci8090173

RICE, S. O. (1944). Mathematical analysis of random noise. *Bell Syst*. *Tech*. *J*. **23** 282–332. MR0010932 https://doi.org/10.1002/j.1538-7305.1944.tb00874.x

ROSEN, B. R. and SAVOY, R. L. (2012). fMRI at 20: Has it changed the world? *NeuroImage* **62** 1316–1324. https://doi.org/10.1016/j.neuroimage.2012.03.004

ROWE, D. B. (2005a). Modeling both the magnitude and phase of complex-valued fMRI data. *NeuroImage* **25** 1310–1324.

ROWE, D. B. (2005b). Parameter estimation in the magnitude-only and complex-valued fMRI data models. *NeuroImage* **25** 1124–1132.

ROWE, D. B. (2016). Handbook of Neuroimaging Data Analysis Image Reconstruction in Functional MRI, 205–232. Chapman & Hall/CRC, London.

ROWE, D. B. and LOGAN, B. R. (2004). A complex way to compute fMRI activation. *NeuroImage* **23** 1078–1092. https://doi.org/10.1016/j.neuroimage.2004.06.042

RUMEAU, C., TZOURIO, N., MURAYAMA, N., PERETTI-VITON, P., LEVRIER, O., JOLIOT, M., MAZOYER, B. and SALAMON, G. (1994). Location of hand function in the sensorimotor cortex: MR and functional correlation. *Am*. *J*. *Neuroradiol*. **15** 567–572.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann*. *Statist*. **6** 461–464. MR0468014

SIJBERS, J., DEN DEKKER, A. J., SCHEUNDERS, P. and DYCK, D. V. (1998). Maximum-likelihood estimation of Rician distribution parameters. *IEEE Trans*. *Med*. *Imag*. **17** 357–361.

SOLO, V. and NOH, J. (2007). An EM algorithm for Rician fMRI activation detection. In *ISBI* 464-467.

TRISTÁN-VEGA, A., AJA-FERNÁNDEZ, S. and WESTIN, C.-F. (2012). Least squares for diffusion tensor estimation revisited: Propagation of uncertainty with Rician and non-Rician signals. *NeuroImage* **59** 4032–4043.

TWEIG, D. B. (1983). The k-trajectory formulation of the NMR imaging process with applications in analysis and synthesis of imaging methods. *Me*. *Phys*. **10** 610–21.

WANG, T. and LEI, T. (1994). Statistical analysis of MR imaging and its applications in image modeling. In *Proceedings of the IEEE International Conference on Image Processing and Neural Networks* **1** 866–870.

WANG, Y., HU, Q., LI, L., FOLEY, A. M. and SRINIVASAN, D. (2019). Approaches to wind power curve modeling: A review and discussion. *Renew*. *Sustain*. *Energy Rev*. **116** 109422.

WEGMANN, B., EKLUND, A. and VILLANI, M. (2017). Bayesian Rician regression for neuroimaging. *Front*. *Neurosci*. **11** 586. https://doi.org/10.3389/fnins.2017.00586

WOO, C.-W., KRISHNAN, A. and WAGER, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage* **91** 412–419. https://doi.org/10.1016/j.neuroimage.2013.12.058

WORSLEY, K. J., MARRETT, S., NEELIN, P., VANDAL, A. C., FRISTON, K. J. and EVANS, A. C. (1996). A unified statistical approach for determining significant voxels in images of cerebral activation. *Hum*. *Brain Mapp*. **4** 58–73.

YU, C.-H., PRADO, R., OMBAO, H. and ROWE, D. (2018). A Bayesian variable selection approach yields improved detection of brain activation from complex-valued fMRI. *J*. *Amer*. *Statist*. *Assoc*. **113** 1395–1410. MR3902218 https://doi.org/10.1080/01621459.2018.1476244

ZHANG, J. L., RUSINEK, H., CHANDARANA, H. and LEE, V. S. (2013). Functional MRI of the kidneys. *J*. *Magn*. *Reson*. *Imaging* **37** 282–293.

ZHOU, X.-H., OBUCHOWSKI, N. A. and MCCLISH, D. K. (2011). *Statistical Methods in Diagnostic Medicine*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2816760 https://doi.org/10.1002/9780470906514

ZHU, H., LI, Y., IBRAHIM, J. G., SHI, X., AN, H., CHEN, Y., GAO, W., LIN, W., ROWE, D. B. et al. (2009). Regression models for identifying noise sources in magnetic resonance images. *J*. *Amer*. *Statist*. *Assoc*. **104** 623–637. MR2751443 https://doi.org/10.1198/jasa.2009.0029

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1514–1532
https://doi.org/10.1214/24-AOAS2000

FUNCTIONAL DATA ANALYSIS WITH MÖBIUS WAVES: APPLICATIONS TO BIOMEDICAL OSCILLATORY SIGNALS

BY ITZIAR FERNÁNDEZ[a], LARRIBA YOLANDA[b], CHRISTIAN CANEDO[c] AND CRISTINA RUEDA[d]

[1]*Department of Statistics and Operations Research, University of Valladolid,* [a]*itziar.fernandez@uva.es,* [b]*yolanda.larriba@uva.es,* [c]*christian.canedo@uva.es,* [d]*cristina.rueda@uva.es*

The Frequency Modulated Möbius (FMM) approach is a contemporary and versatile tool for analyzing oscillatory signals. Within the functional data analysis framework, the FMM model offers an accurate alternative for comprehending multidimensional synchronized oscillatory signals, which are common in various fields including biology and medicine. This approach decomposes the signals into scaled Möbius waves formulated in terms of four parameters, interpreted as measures of location and shape variation. Among these parameters, two are specific to individual signals, while the remaining two are global and describe the interconnections between signals. In this paper we extend the utility of the FMM by developing inferential procedures based on likelihood estimations for nonlinear models. These procedures cover the core parameters, the signals themselves, and their derivatives. We explore various techniques, including the Gauss–Seidel approach and profile likelihood, for obtaining estimators and their asymptotic properties. Our methodology undergoes rigorous validation, drawing on a combination of theoretical insights and numerical experiments. Furthermore, as our research is primarily motivated by the analysis of biomedical oscillatory signals, we apply this methodology to address two pertinent real-world problems involving electrocardiogram and pattern-reversal visual evoked potential data. These applications illustrate the practical efficacy of the new paradigm.

REFERENCES

CANEDO, C., FERNÁNDEZ, I., COCO, R. M., CUADRADO, R. and RUEDA, C. (2023). Novel modeling proposals for the analysis of pattern electroretinogram signals. In *Statistical Methods at the Forefront of Biomedical Advances* 255–273. Springer.

CARROLL, C., MÜLLER, H.-G. and KNEIP, A. (2021). Cross-component registration for multivariate functional data, with application to growth curves. *Biometrics* **77** 839–851. MR4320661 https://doi.org/10.1111/biom.13340

CHIOU, J.-M. and MÜLLER, H.-G. (2016). A pairwise interaction model for multivariate functional and longitudinal data. *Biometrika* **103** 377–396. MR3509893 https://doi.org/10.1093/biomet/asw007

CLARK, E. N., KATIBI, I. and MACFARLANE, P. W. (2014). Automatic detection of end QRS notching or slurring. *J. Electrocardiol*. **47** 151–154. https://doi.org/10.1016/j.jelectrocard.2013.10.007

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

DE LUNA, A. B. (2008). *Basic Electrocardiography*: *Normal and Abnormal ECG Patterns*. Wiley.

FERNÁNDEZ, I., YOLANDA, L., CANEDO, C. and RUEDA, C. (2025). Supplement to "Functional data analysis with Möbius waves: Applications to biomedical oscillatory signals." https://doi.org/10.1214/24-AOAS2000SUPPA, https://doi.org/10.1214/24-AOAS2000SUPPB

GRITZALI, F. (1988). Towards a generalized scheme for QRS detection in ECG waveforms. *Signal Process*. **15** 183–192. https://doi.org/10.1016/0165-1684(88)90069-2

GROVER, R., KUNDU, D. and MITRA, A. (2021). An efficient methodology to estimate the parameters of a two-dimensional chirp signal model. *Multidimens*. *Syst*. *Signal Process*. **32** 49–75. MR4209899 https://doi.org/10.1007/s11045-020-00728-x

Key words and phrases. FMM model, maximum likelihood estimation, nonlinear model, oscillatory signal, signal decomposition.

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

HOROWITZ, J., KLEMELÄ, J. and MAMMEN, E. (2006). Optimal estimation in additive regression models. *Bernoulli* **12** 271–298. MR2218556 https://doi.org/10.3150/bj/1145993975

HOU, H., MOGHIMI, S., ZANGWILL, L. M., SHOJI, T., GHAHARI, E., MANALASTAS, P. I. C., PENTEADO, R. C. and WEINREB, R. N. (2018). Inter-eye asymmetry of optical coherence tomography angiography vessel density in bilateral glaucoma, glaucoma suspect, and healthy eyes. *Am. J. Ophthalmol.* **190** 69–77. https://doi.org/10.1016/j.ajo.2018.03.026

HUANG-LINK, Y., YANG, G., GUSTAFSSON, G., GAUFFIN, H., LANDTBLOM, A.-M., MIRABELLI, P. and LINK, H. (2023). The importance of optical coherence tomography in the diagnosis of atypical or subclinical optic neuritis: A case series study. *J. Clin. Med.* **12** 1309.

JAGADAMBA, A. and KARTHIYANEE, K. (2012). Effect of ocular dominance on the latency and amplitude of visual evoked potentials. *IOSR J. Dent. Med. Sci.* **2** 19–24.

KOTHARI, R., BOKARIYA, P., SINGH, S. and SINGH, R. (2016). A comprehensive review on methodologies employed for visual evoked potentials. *Scientifica* **2016** 9852194. https://doi.org/10.1155/2016/9852194

KURITA-TASHIMA, S., TOBIMATSU, S., NAKAYAMA-HIROMATSU, M. and KATO, M. (1991). Effect of check size on the pattern reversal visual evoked potential. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section* **80** 161–166.

MERDJANOVSKA, E. and RASHKOVSKA, A. (2022). Comprehensive survey of computational ECG analysis: Databases, methods and applications. *Expert Syst. Appl.* 117206. https://doi.org/10.1016/j.eswa.2022.117206

MONEDERO, I. (2022). A novel ECG diagnostic system for the detection of 13 different diseases. *Engineering Applications of Artificial Intelligence* **107** 104536. https://doi.org/10.1016/j.engappai.2021.104536

MUTHUSWAMY, J. (2003). Biomedical signal analysis. In *Standard Handbook of Biomedical Engineering and Design* McGraw-Hill.

NETO, S. P., PINTO, L. C. and ALVARENGA, R. M. P. (2017). The visual evoked potential in idiopathic inflammatory demyelinating diseases. In *Event-Related Potentials and Evoked Potentials* InTech Rijeka.

ODOM, J. V., BACH, M., BRIGELL, M., HOLDER, G. E., MCCULLOCH, D. L., MIZOTA, A., TORMENE, A. P. and INTERNATIONAL SOCIETY FOR CLINICAL ELECTROPHYSIOLOGY OF VISION (2016). ISCEV standard for clinical visual evoked potentials: (2016 update). *Doc. Ophthalmol.* **133** 1–9.

PAPADOPOULOU, A., PFISTER, A., TSAGKAS, C., GAETANO, L., SELLATHURAI, S., D'SOUZA, M., CERDÁ-FUERTES, N., GUGLETA, K., DESCOTEAUX, M. et al. (2024). Visual evoked potentials in multiple sclerosis: P100 latency and visual pathway damage including the lateral geniculate nucleus. *Clin. Neurophysiol.* **161** 122–132. https://doi.org/10.1016/j.clinph.2024.02.020

PARISI, V., SCARALE, M. E., BALDUCCI, N., FRESINA, M. and CAMPOS, E. C. (2010). Electrophysiological detection of delayed postretinal neural conduction in human amblyopia. *Investig. Ophthalmol. Vis. Sci.* **51** 5041–5048.

PASSAMONTI, L., CERASA, A., LIGUORI, M., GIOIA, M. C., VALENTINO, P., NISTICÒ, R., QUATTRONE, A. and FERA, F. (2009). Neurobiological mechanisms underlying emotional processing in relapsing-remitting multiple sclerosis. *Brain* **132** 3380–3391. https://doi.org/10.1093/brain/awp095

PERERA, N. D. and DALUWATTE, C. (2018). Detecting strict left bundle branch block from 12-lead electrocardiogram using support vector machine classification and derivative analysis. In 2018 *Computing in Cardiology Conference* (*CinC*) **45** 1–4. https://doi.org/10.22489/CinC.2018.030

QIAO, X., QIAN, C., JAMES, G. M. and GUO, S. (2020). Doubly functional graphical models in high dimensions. *Biometrika* **107** 415–431. MR4108937 https://doi.org/10.1093/biomet/asz072

ROSS, G. (1990). *Nonlinear Estimation*. Springer, New York.

RUEDA, C., FERNÁNDEZ, I., LARRIBA, Y. and RODRÍGUEZ-COLLADO, A. (2021). The FMM approach to analyze biomedical signals: Theory, software, applications and future. *Mathematics* **9** 1145. https://doi.org/10.3390/math9101145

RUEDA, C., LARRIBA, Y. and LAMELA, A. (2021). The hidden waves in the ECG uncovered revealing a sound automated interpretation method. *Sci. Rep.* **11** 1–11. https://doi.org/10.1038/s41598-021-82520-w

RUEDA, C., LARRIBA, Y. and PEDDADA, S. D. (2019). Frequency modulated Möbius model accurately predicts rhythmic signals in biological and physical sciences. *Sci. Rep.* **9** 18701. https://doi.org/10.1038/s41598-019-54569-1

RUEDA, C., RODRÍGUEZ-COLLADO, A., FERNÁNDEZ, I., CANEDO, C., UGARTE, M. D. and LARRIBA, Y. (2022). A unique cardiac electrocardiographic 3D model. Toward interpretable AI diagnosis. *iScience* **25** 105617. https://doi.org/10.1016/j.isci.2022.105617

RUEDA, C., RODRÍGUEZ-COLLADO, A. and LARRIBA, Y. (2021). A novel wave decomposition for oscillatory signals. *IEEE Trans. Signal Process.* **69** 960–972. MR4213393 https://doi.org/10.1109/TSP.2021.3051428

SEBER, G. A. F. and WILD, C. J. (2003). *Nonlinear Regression*. Wiley, New York. MR0986070 https://doi.org/10.1002/0471725315

SMISEK, R., VISCOR, I., JURAK, P., HALAMEK, J. and PLESINGER, F. (2018). Fully automatic detection of strict left bundle branch block. *J. Electrocardiol*. **51** S31–S34. https://doi.org/10.1016/j.jelectrocard.2018.06.013

TOOSY, A. T., MASON, D. F. and MILLER, D. H. (2014). Optic neuritis. *Lancet Neurol*. **13** 83–99. https://doi.org/10.1016/S1474-4422(13)70259-X

VIDAL-JORDANA, A., ROVIRA, A., ARRAMBIDE, G., OTERO-ROMERO, S., RÍO, J., COMABELLA, M., NOS, C., CASTILLÓ, J., GALAN, I. et al. (2021). Optic nerve topography in multiple sclerosis diagnosis: The utility of visual evoked potentials. *Neurology* **96** e482–e490.

WAGNER, P., STRODTHOFF, N., BOUSSELJOT, R.-D., KREISELER, D., LUNZE, F. I., SAMEK, W. and SCHAEFFTER, T. (2020). PTB-XL, a large publicly available electrocardiography dataset. *Sci*. *Data* **7** 1–15. https://doi.org/10.6084/m9.figshare.12098055

YANG, L. and TANG, X. (2019). Bilateral asymmetry improved accuracy when assessing glaucomatous vision-related quality of life impairment. *Medicine* **98** e17924.

ZHOU, L., HUANG, J. Z. and CARROLL, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika* **95** 601–619. MR2443178 https://doi.org/10.1093/biomet/asn035

ZONG, W., MOODY, G. and JIANG, D. (2003). A robust open-source algorithm to detect onset and duration of QRS complexes. In *Computers in Cardiology*, 2003 737–740. IEEE. https://doi.org/10.1109/CIC.2003.1291261

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1533–1552
https://doi.org/10.1214/24-AOAS2010

A GENERAL FRAMEWORK OF BRAIN REGION DETECTION AND GENETIC VARIANTS SELECTION IN IMAGING GENETICS

BY SIQIANG SU[1,a], ZHENGHAO LI[1,b], LONG FENG[1,c] AND TING LI[2,d]

[1] *Department of Statistics & Actuarial Science, School of Computing & Data Science, University of Hong Kong,* [a] *u3008662@connect.hku.hk,* [b] *lizh@connect.hku.hk,* [c] *lfeng@hku.hk*

[2] *Department of Applied Mathematics, Hong Kong Polytechnic University,* [d] *tingeric.li@polyu.edu.hk*

Imaging genetics is a growing field that employs structural or functional neuroimaging techniques to study individuals with genetic risk variants potentially linked to specific illnesses. This area presents considerable challenges to statisticians due to the heterogeneous information and different data forms it involves. In addition, both imaging and genetic data are typically high-dimensional, creating a "big data squared" problem, Moreover, brain imaging data contains extensive spatial information. Simply vectorizing tensor images and treating voxels as independent features can lead to computational issues and disregard spatial structure. This paper presents a novel statistical method for imaging genetics modeling while addressing all these challenges. We explore a canonical correlation analysis based linear model for the joint modeling of brain imaging, genetic information, and clinical phenotype, enabling the simultaneous detection of significant brain regions and selection of important genetic variants associated with the phenotype outcome. Scalable algorithms are developed to tackle the "big data squared" issue. We apply the proposed method to explore the reaction speed, an indicator of cognitive functions, and its associations with brain MRI and genetic factors using the UK Biobank database. Our study reveals a notable connection between the caudate nucleus region of brain and specific significant SNPs, along with their respective regulated genes, and the reaction speed.

REFERENCES

BROUWER, R. M., KLEIN, M., GRASBY, K. L., SCHNACK, H. G., JAHANSHAD, N., TEEUW, J., THOMOPOULOS, S. I., SPROOTEN, E., FRANZ, C. E. et al. (2022). Genetic variants associated with longitudinal changes in brain structure across the lifespan. *Nat. Neurosci*. **25** 421–432.

BYCROFT, C., FREEMAN, C. and PETKOVA, D. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562** 203–209.

CAI, N., REVEZ, J. A., ADAMS, M. J., ANDLAUER, T. F., BREEN, G., BYRNE, E. M., CLARKE, T.-K., FORSTNER, A. J., GRABE, H. J. et al. (2020). Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat. Genet*. **52** 437–447.

CHI, E. C., ALLEN, G. I., ZHOU, H., KOHANNIM, O., LANGE, K. and THOMPSON, P. M. (2013). Imaging genetics via sparse canonical correlation analysis. In 2013 *IEEE* 10*th International Symposium on Biomedical Imaging* 740–743. IEEE Press, New York.

DAVIES, G., LAM, M., HARRIS, S. E., TRAMPUSH, J. W., LUCIANO, M., HILL, W. D., HAGENAARS, S. P., RITCHIE, S. J., MARIONI, R. E. et al. (2018). Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat. Commun*. **9**. 2098.

DAVIES, G., MARIONI, R. E., LIEWALD, D. C., HILL, W. D., HAGENAARS, S. P., HARRIS, S. E., RITCHIE, S. J., LUCIANO, M., FAWNS-RITCHIE, C. et al. (2016). Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N = 112,151). *Mol. Psychiatry* **21** 758–767.

DOORNIK, J. A. and HANSEN, H. (1994). Non-linear registration aka spatial normalisation. Internal Technical Report TR07JA2, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Dept. Clinical Neurology, Oxford Univ.

DRISCOLL, M. E., BOLLU, P. C. and TADI, P. (2020). Neuroanatomy, nucleus caudate.

Key words and phrases. Neuroimaging data, imaging genetics, cognitive functions, canonical correlation analysis, Kronecker product.

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

GE, T., FENG, J., HIBAR, D. P., THOMPSON, P. M. and NICHOLS, T. E. (2012). Increasing power for voxel-wise genome-wide association studies: The random field theory, least square kernel machines and fast permutation procedures. *NeuroImage* **63** 858–873. https://doi.org/10.1016/j.neuroimage.2012.07.012

GIANNAKOPOULOU, O., LIN, K., MENG, X., SU, M.-H., KUO, P.-H., PETERSON, R. E., AWASTHI, S., MOSCATI, A., COLEMAN, J. R. et al. (2021). The genetic architecture of depression in individuals of East Asian ancestry: A genome-wide association study. *JAMA Psychiatr.* **78** 1258–1269.

GIRKA, F., GLOAGUEN, A., LE BRUSQUET, L., ZUJOVIC, V. and TENENHAUS, A. (2024). Tensor generalized canonical correlation analysis. *Inf. Fusion* **102** 102045.

GOLDSMITH, J., HUANG, L. and CRAINICEANU, C. M. (2014). Smooth scalar-on-image regression via spatial Bayesian variable selection. *J. Comput. Graph. Statist.* **23** 46–64. MR3173760 https://doi.org/10.1080/10618600.2012.743437

GRABNER, G., JANKE, A. L., BUDGE, M. M., SMITH, D., PRUESSNER, J. and COLLINS, D. L. (2006). Symmetric atlasing and model based segmentation: An application to the hippocampus in older adults. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI* 2006 58–66. Springer, Berlin Heidelberg.

GRAHN, J. A., PARKINSON, J. A. and OWEN, A. M. (2008a). The cognitive functions of the caudate nucleus. *Prog. Neurobiol.* **86** 141–155.

GRAHN, J. A., PARKINSON, J. A. and OWEN, A. M. (2008b). The cognitive functions of the caudate nucleus. *Prog. Neurobiol.* **86** 141–155.

GRAHN, J. A., PARKINSON, J. A. and OWEN, A. M. (2009). The role of the basal ganglia in learning and memory: Neuropsychological studies. *Behav. Brain Res.* **199** 53–60. https://doi.org/10.1016/j.bbr.2008.11.020

HANAFI, M. and KIERS, H. A. L. (2006). Analysis of K sets of data, with differential emphasis on agreement between and within sets. *Comput. Statist. Data Anal.* **51** 1491–1508. MR2307521 https://doi.org/10.1016/j.csda.2006.04.020

HAWI, Z., YATES, H., PINAR, A., ARNATKEVICIUTE, A., JOHNSON, B., TONG, J., PUGSLEY, K., DARK, C., PAUPER, M. et al. (2018). A case–control genome-wide association study of ADHD discovers a novel association with the tenascin R (TNR) gene. *Transl. Psychiatry* **8** 284.

HUANG, M., NICHOLS, T., HUANG, C., YU, Y., LU, Z., KNICKMEYER, R. C., FENG, Q., ZHU, H., ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE et al. (2015). FVGWAS: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *NeuroImage* **118** 613–627.

INTERNATIONAL MULTIPLE SCLEROSIS GENETICS CONSORTIUM, ANZGENE, IIBDGC and WTCCC2 (2019). Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365** eaav7188.

JENKINSON, M., BANNISTER, P., BRADY, M. and SMITH, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* **17** 825–841. https://doi.org/10.1016/s1053-8119(02)91132-8

JIJI, S., SMITHA, K. A., GUPTA, A. K., PILLAI, V. P. M. and JAYASREE, R. S. (2013). Segmentation and volumetric analysis of the caudate nucleus in Alzheimer's disease. *Eur. J. Radiol.* **82** 1525–1530. https://doi.org/10.1016/j.ejrad.2013.03.012

KANG, J., REICH, B. J. and STAICU, A.-M. (2018). Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika* **105** 165–184. MR3768872 https://doi.org/10.1093/biomet/asx075

KENDALL, K. M., BRACHER-SMITH, M., FITZPATRICK, H., LYNHAM, A., REES, E., ESCOTT-PRICE, V., OWEN, M. J., O'DONOVAN, M. C., WALTERS, J. T. R. et al. (2019). Cognitive performance and functional outcomes of carriers of pathogenic copy number variants: Analysis of the UK Biobank. *Br. J. Psychiatry* **214** 297–304. https://doi.org/10.1192/bjp.2018.301

KETTENRING, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika* **58** 433–451. MR0341750 https://doi.org/10.1093/biomet/58.3.433

KNUTSON, B., ADAMS, C. M., FONG, G. W. and HOMMER, D. (2001a). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J. Neurosci.* **21**. RC159.

KNUTSON, B., FONG, G. W., ADAMS, C. M., VARNER, J. L. and HOMMER, D. (2001b). Dissociation of reward anticipation and outcome with event-related fMRI. *NeuroReport* **12** 3683–3687.

LANGE, K., WON, J.-H., LANDEROS, A. and ZHOU, H. (2021). Nonconvex optimization via MM algorithms: Convergence theory. In *Computational Statistics in Data Science* 1–22. Wiley, New York. https://doi.org/10.1002/9781118445112.stat08295

LATOURELLE, J. C., PANKRATZ, N., DUMITRIU, A., WILK, J. B., GOLDWURM, S., PEZZOLI, G., MARIANI, C. B., DESTEFANO, A. L., HALTER, C. et al. (2009). Genomewide association study for onset age in Parkinson disease. *BMC Med. Genet.* **10** 1–14.

LEMHADRI, I., RUAN, F., ABRAHAM, L. and TIBSHIRANI, R. (2021). LassoNet: A neural network with feature sparsity. *J. Mach. Learn. Res.* **22** Paper No. 127, 29. MR4279778

LI, T., HU, J., WANG, S. and ZHANG, H. (2021). Super-variants identification for brain connectivity. *Hum. Brain Mapp.* **42** 1304–1312.

LIU, Z., YANG, N., DONG, J., TIAN, W., CHANG, L., MA, J., GUO, J., TAN, J., DONG, A. et al. (2022). Deficiency in endocannabinoid synthase DAGLB contributes to early onset Parkinsonism and murine nigral dopaminergic neuron dysfunction. *Nat. Commun.* **13** 3490.

LUAN, Z., ZHANG, Y., LU, T., RUAN, Y., ZHANG, H., YAN, J., LI, L., SUN, W., WANG, L. et al. (2011). Positive association of the human STON2 gene with schizophrenia. *NeuroReport* **22** 288–293.

LUCIANO, M., HAGENAARS, S. P., DAVIES, G., HILL, W. D., CLARKE, T.-K., SHIRALI, M., HARRIS, S. E., MARIONI, R. E., LIEWALD, D. C. et al. (2018). Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nat. Genet.* **50** 6–11.

LYALL, D. M., CELIS-MORALES, C. A., ANDERSON, J., GILL, J. M. R., MACKAY, D. F., MCINTOSH, A. M., SMITH, D. J., DEARY, I. J., SATTAR, N. et al. (2017). Associations between single and multiple cardiometabolic diseases and cognitive abilities in 474 129 UK Biobank participants. *Eur. Heart J.* **38** 577–583. https://doi.org/10.1093/eurheartj/ehw528

LYALL, D. M., QUINN, T., LYALL, L. M., WARD, J., ANDERSON, J. J., SMITH, D. J., STEWART, W., STRAWBRIDGE, R. J., BAILEY, M. E. S. et al. (2022). Quantifying bias in psychological and physical health in the UK Biobank imaging sub-sample. *Brain Commun.* **4** fcac119. https://doi.org/10.1093/braincomms/fcac119

MIN, E. J., CHI, E. C. and ZHOU, H. (2019). Tensor canonical correlation analysis. *Stat* **8** e253, 11. MR4072118 https://doi.org/10.1002/sta4.253

NATHOO, F. S., KONG, L., ZHU, H. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2019). A review of statistical methods in imaging genetics. *Canad. J. Statist.* **47** 108–131. MR3919897 https://doi.org/10.1002/cjs.11487

OERTEL-KNÖCHEL, V., LANCASTER, T. M., KNÖCHEL, C., STÄBLEIN, M., STORCHAK, H., REINKE, B., JURCOANE, A., KNIEP, J., PRVULOVIC, D. et al. (2015). Schizophrenia risk variants modulate white matter volume across the psychosis spectrum: Evidence from two independent cohorts. *NeuroImage Clin.* **7** 764–770.

ROVIRA, P., DEMONTIS, D., SÁNCHEZ-MORA, C., ZAYATS, T., KLEIN, M., MOTA, N. R., WEBER, H., GARCIA-MARTÍNEZ, I., PAGEROLS, M. et al. (2020). Shared genetic background between children and adults with attention deficit/hyperactivity disorder. *Neuropsychopharmacology* **45** 1617–1626.

RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D, Nonlinear Phenom.* **60** 259–268. MR3363401 https://doi.org/10.1016/0167-2789(92)90242-F

INTERNATIONAL MULTIPLE SCLEROSIS GENETICS CONSORTIUM, WELLCOME TRUST CASE CONTROL CONSORTIUM 2, SAWCER, S., HELLENTHAL, G., PIRINEN, M., SPENCER, C. C., PATSOPOULOS, N. A., MOUTSIANAS, L., DILTHEY, A. et al. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476** 214–219.

SCHRIMSHER, G. W., BILLINGSLEY, R. L., JACKSON, E. F. and MOORE, B. D. (2002). Caudate nucleus volume asymmetry predicts attention-deficit hyperactivity disorder (ADHD) symptomatology in children. *J. Child. Neurol.* **17** 877–884. https://doi.org/10.1177/08830738020170122001

SMITH, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* **17** 143–155. https://doi.org/10.1002/hbm.10062

STEIN, J. L., HUA, X., LEE, S., HO, A. J., LEOW, A. D., TOGA, A. W., SAYKIN, A. J., SHEN, L., FOROUD, T. et al. (2010). Voxelwise genome-wide association study (vGWAS). *NeuroImage* **53** 1160–1174.

SU, S., LI, Z., FENG, L. and LI, T. (2025). Supplement to "A General Framework of Brain Region Detection and Genetic Variants Selection in Imaging Genetics." https://doi.org/10.1214/24-AOAS2010SUPPA, https://doi.org/10.1214/24-AOAS2010SUPPB

TAKASE, K., TAMAGAKI, C., OKUGAWA, G., NOBUHARA, K., MINAMI, T., SUGIMOTO, T., SAWADA, S. and KINOSHITA, T. (2004). Reduced white matter volume of the caudate nucleus in patients with schizophrenia. *Neuropsychobiology* **50** 296–300. https://doi.org/10.1159/000080956

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

TRUBETSKOY, V., PARDIÑAS, A. F., QI, T., PANAGIOTAROPOULOU, G., AWASTHI, S., BIGDELI, T. B., BRYOIS, J., CHEN, C.-Y., DENNISON, C. A. et al. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604** 502–508.

TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* **109** 475–494. MR1835069 https://doi.org/10.1023/A:1017501703105

VAN DE GEER, J. P. (1984). Linear relations among k sets of variables. *Psychometrika* **49** 79–94.

VOUNOU, M., NICHOLS, T. E., MONTANA, G., ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE et al. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. *NeuroImage* **53** 1147–1159.

WANG, H., NIE, F., HUANG, H., KIM, S., NHO, K., RISACHER, S. L., SAYKIN, A. J., SHEN, L. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2012). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics* **28** 229–237.

WANG, X., ZHU, H. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2017). Generalized scalar-on-image regression models via total variation. *J. Amer. Statist. Assoc.* **112** 1156–1168. MR3735367 https://doi.org/10.1080/01621459.2016.1194846

WU, S. and FENG, L. (2023). Sparse Kronecker product decomposition: A general framework of signal region detection in image regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 783–809. MR4745944 https://doi.org/10.1093/jrsssb/qkad024

XU, Y. and YIN, W. (2013). A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM J. Imaging Sci.* **6** 1758–1789. MR3105787 https://doi.org/10.1137/120887795

YU, D., WANG, L., KONG, D. and ZHU, H. (2022). Mapping the genetic-imaging-clinical pathway with applications to Alzheimer's disease. *J. Amer. Statist. Assoc.* **117** 1656–1668. MR4528461 https://doi.org/10.1080/01621459.2022.2087658

ZHAO, B., LI, T., FAN, Z., YANG, Y., SHU, J., YANG, X., WANG, X., LUO, T., TANG, J. et al. (2023a). Heart–brain connections: Phenotypic and genetic insights from magnetic resonance images. *Science* **380**. abn6598.

ZHAO, B., LI, T., FAN, Z., YANG, Y., SHU, J., YANG, X., WANG, X., LUO, T., TANG, J. et al. (2023b). Heart–brain connections: Phenotypic and genetic insights from magnetic resonance images. *Science* **380**.

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**.

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

ZHU, H., LI, T. and ZHAO, B. (2023). Statistical learning methods for neuroimaging data analysis with applications. *Annu. Rev. Biomed. Data Sci.* **6** 73–104.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1553–1577
https://doi.org/10.1214/25-AOAS2022

FUNCTIONAL FACTOR MODELING OF BRAIN CONNECTIVITY

BY KYLE STANLEY[1,a], NICOLE LAZAR[1,2,b] AND MATTHEW REIMHERR[1,3,c]

[1]*Department of Statistics, Pennsylvania State University,* [a]*kms8227@psu.edu*
[2]*Huck Institutes of the Life Sciences, Pennsylvania State University,* [b]*nfl5182@psu.edu*
[3]*Amazon Science,* [c]*mlr36@psu.edu*

Many fMRI analyses examine functional connectivity, or statistical dependencies among remote brain regions. Factor analysis, which parsimoniously describes correlations between many observed variables, offers a natural framework in which to study such dependencies. However, multivariate factor models break down when applied to functional and spatiotemporal data, like fMRI. We present a factor model for discretely-observed multidimensional functional data that is well suited to the study of functional connectivity. Unlike classical factor models which decompose a multivariate observation into a "common" term that captures covariance between observed variables and an uncorrelated "idiosyncratic" term that captures variance unique to each observed variable, our model decomposes a functional observation into two uncorrelated components: a "global" term that captures long-range dependencies and a "local" term that captures short-range dependencies. We show that if the global covariance is smooth with finite rank and the local covariance is banded with potentially infinite rank, then this decomposition is identifiable. Under these conditions, recovery of the global covariance amounts to rank-constrained matrix completion, which we exploit to formulate consistent estimators. Through simulations and an application to resting-state fMRI data, we demonstrate that our approach offers several distinct advantages over popular functional connectivity methods.

REFERENCES

AFYOUNI, S., SMITH, S. M. and NICHOLS, T. E. (2019). Effective degrees of freedom of the Pearson's correlation coefficient under autocorrelation. *NeuroImage* **199** 609–625. https://doi.org/10.1016/j.neuroimage.2019.05.011

BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171. MR1956857 https://doi.org/10.1111/1468-0262.00392

BECKMANN, C. F. and SMITH, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans*. *Med*. *Imag*. **23** 137–152.

BISWAL, B., ZERRIN YETKIN, F., HAUGHTON, V. M. and HYDE, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn*. *Reson*. *Med*. **34** 537–541.

BROYDEN, C. G. (1970). The convergence of a class of double-rank minimization algorithms. I. General considerations. *IMA J*. *Appl*. *Math*. **6** 76–90.

BULLMORE, E., LONG, C., SUCKLING, J., FADILI, J., CALVERT, G., ZELAYA, F., CARPENTER, T. A. and BRAMMER, M. (2001). Colored noise and computational inference in neurophysiological (fMRI) time series analysis: Resampling methods in time and wavelet domains. *Hum*. *Brain Mapp*. **12** 61–78. https://doi.org/10.1002/1097-0193(200102)12:2<61::aid-hbm1004>3.0.co;2-w

CALHOUN, V. D., ADALI, T., PEARLSON, G. D. and PEKAR, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Hum*. *Brain Mapp*. **14** 140–151. https://doi.org/10.1002/hbm.1048

CHAMBERLAIN, G. and ROTHSCHILD, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** 1281–1304. MR0736050 https://doi.org/10.2307/1912275

CHRISTOVA, P., LEWIS, S. M., JERDE, T. A., LYNCH, J. K. and GEORGOPOULOS, A. P. (2011). True associations between resting fMRI time series based on innovations. *J*. *Neural Eng*. **8** 046025. https://doi.org/10.1088/1741-2560/8/4/046025

Key words and phrases. Functional data analysis, factor analysis, fMRI, functional connectivity, matrix completion.

DESCARY, M.-H. and PANARETOS, V. M. (2019). Functional data analysis by matrix completion. *Ann. Statist.* **47** 1–38. MR3909925 https://doi.org/10.1214/17-AOS1590

ESTEBAN, O., MARKIEWICZ, C. J., BLAIR, R. W., MOODIE, C. A., ISIK, A. I., ERRAMUZPE, A., KENT, J. D., GONCALVES, M., DUPRE, E. et al. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nat. Methods* **16** 111–116.

FLETCHER, R. (1970). A new approach to variable metric algorithms. *Comput. J.* **13** 317–322.

FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Rev. Econ. Stat.* **82** 540–554. https://doi.org/10.1162/003465300559037

FRISTON, K. J. (1998). Modes or models: A critique on independent component analysis for fMRI. *Trends Cogn. Sci.* **2** 373–375. https://doi.org/10.1016/s1364-6613(98)01227-3

FRISTON, K. J. (2011). Functional and effective connectivity: A review. *Brain Connect.* **1** 13–36. https://doi.org/10.1089/brain.2011.0008

FRISTON, K. J., FRITH, C. D., LIDDLE, P. F. and FRACKOWIAK, R. S. (1993). Functional connectivity: The principal-component analysis of large (PET) data sets. *J. Cereb. Blood Flow Metab.* **13** 5–14.

GEWEKE, J. (1977). The dynamic factor analysis of economic time series. *Latent Var. Socio-Econ. Models* 365–383.

GOLDFARB, D. (1970). A family of variable-metric methods derived by variational means. *Math. Comp.* **24** 23–26. MR0258249 https://doi.org/10.2307/2004873

HAYS, S., SHEN, H. and HUANG, J. Z. (2012). Functional dynamic factor models with application to yield curve forecasting. *Ann. Appl. Stat.* **6** 870–894. MR3012513 https://doi.org/10.1214/12-AOAS551

HÖRMANN, S. and JAMMOUL, F. (2022). Preprocessing noisy functional data: A multivariate perspective. *Electron. J. Stat.* **16** 6232–6266. MR4515715 https://doi.org/10.1214/22-ejs2083

HYNDMAN, R. J. and KHANDAKAR, Y. (2008). Automatic time series forecasting: The forecast package for R. *J. Stat. Softw.* **27** 1–22.

HYVÄRINEN, A. and OJA, E. (2000). Independent component analysis: Algorithms and applications. *Neural Netw.* **13** 411–430.

IRAJI, A., FU, Z., FAGHIRI, A., DUDA, M., CHEN, J., RACHAKONDA, S., DERAMUS, T., KOCHUNOV, P., ADHIKARI, B. et al. (2023). Identifying canonical and replicable multi-scale intrinsic connectivity networks in 100k+ resting-state fMRI datasets. *Hum. Brain Mapp.* **44** 5729–5748.

JENKINSON, M., BECKMANN, C. F., BEHRENS, T. E. J., WOOLRICH, M. W. and SMITH, S. M. (2012). FSL. *NeuroImage* **62** 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

JOHNSON, W. A. and WICHERN, D. W. (2002). *Applied Multivariate Statistical Analysis*, Prentice Hall, Upper Saddle River.

KAISER, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23** 187–200.

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

KOWAL, D. R., MATTESON, D. S. and RUPPERT, D. (2017). A Bayesian multivariate functional dynamic linear model. *J. Amer. Statist. Assoc.* **112** 733–744. MR3671766 https://doi.org/10.1080/01621459.2016.1165104

LAIRD, A. R., ROGERS, B. P. and MEYERAND, M. E. (2004). Comparison of Fourier and wavelet resampling methods. *Magn. Reson. Med.* **51** 418–422.

LI, B. and SOLEA, E. (2018). A nonparametric graphical model for functional data with application to brain networks based on fMRI. *J. Amer. Statist. Assoc.* **113** 1637–1655. MR3902235 https://doi.org/10.1080/01621459.2017.1356726

LIEBL, D. (2013). Modeling and forecasting electricity spot prices: A functional data perspective. *Ann. Appl. Stat.* **7** 1562–1592. MR3127959 https://doi.org/10.1214/13-AOAS652

LINDQUIST, M. A. (2008). The statistical analysis of fMRI data. *Statist. Sci.* **23** 439–464. MR2530545 https://doi.org/10.1214/09-STS282

LIU, D. C. and NOCEDAL, J. (1989). On the limited memory BFGS method for large scale optimization. *Math. Program.* **45** 503–528. MR1038245 https://doi.org/10.1007/BF01589116

NEUHAUS, J. O. and WRIGLEY, C. (1954). The quartimax method: An analytic approach to orthogonal simple structure 1. *Br. J. Math. Stat. Psychol.* **7** 81–91.

OTTO, S. and SALISH, N. (2022). Dynamic Factor Model for Functional Time Series: Identification, Estimation, and Prediction.

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

REISS, P. T., HUO, L., ZHAO, Y., KELLY, C. and OGDEN, R. T. (2015). Wavelet-domain regression and predictive inference in psychiatric neuroimaging. *Ann. Appl. Stat.* **9** 1076–1101. MR3371348 https://doi.org/10.1214/15-AOAS829

SARGENT, T. and SIMS, C. (1977). Business cycle modeling without pretending to have too much a priori economic theory. Working Papers No. 55. Federal Reserve Bank of Minneapolis.

SHANNO, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Math. Comp.* **24** 647–656. MR0274029 https://doi.org/10.2307/2004840

SMITH, S. M., FOX, P. T., MILLER, K. L., GLAHN, D. C., FOX, P. M., MACKAY, C. E., FILIPPINI, N., WATKINS, K. E., TORO, R. et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. USA* **106** 13040–13045.

SMITH, S. M., HYVÄRINEN, A., VAROQUAUX, G., MILLER, K. L. and BECKMANN, C. F. (2014). Group-PCA for very large fMRI datasets. *NeuroImage* **101** 738–749. https://doi.org/10.1016/j.neuroimage.2014.07.051

SNOEK, L., VAN DER MIESEN, M. M., BEEMSTERBOER, T., VAN DER LEIJ, A., EIGENHUIS, A. and STEVEN SCHOLTE, H. (2021). The Amsterdam open MRI collection, a set of multimodal MRI datasets for individual difference analyses. *Sci. Data* **8** 1–23.

STANLEY, K., LAZAR, N. and REIMHERR, M. (2025). Supplement to "Functional Factor Modeling of Brain Connectivity." https://doi.org/10.1214/25-AOAS2022SUPPA, https://doi.org/10.1214/25-AOAS2022SUPPB

STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179. MR1951271 https://doi.org/10.1198/016214502388618960

VAN DEN HEUVEL, M. P. and POL, H. E. H. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.* **20** 519–534.

VIVIANI, R., GRÖN, G. and SPITZER, M. (2005). Functional principal component analysis of fMRI data. *Hum. Brain Mapp.* **24** 109–129. https://doi.org/10.1002/hbm.20074

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

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1578–1602
https://doi.org/10.1214/25-AOAS2026

COOPERATIVE DIFFERENTIAL NETWORK LEARNING WITH HUB DETECTION FOR MULTICENTER NEUROIMAGING DATA

BY HAO CHEN[1,a], DINGZI GUO[2,b], YING GUO[3,d], YONG HE[2,c], DONG LIU[4,e], LEI LIU[5,f], YUE YIN[6,g] AND XIAO-HUA ZHOU[7,h]

[1]*Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University,* [a]*hao.chen@sdu.edu.cn*

[2]*Zhongtai Securities Institute for Financial Studies, Shandong University ,* [b]*guodingzi@mail.sdu.edu.cn,* [c]*heyong@sdu.edu.cn*

[3]*Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University,* [d]*yguo2@emory.edu*

[4]*School of Physical and Mathematical Sciences, Nanyang Technological University,* [e]*liudong_stat@163.com*

[5]*Division of Biostatistics, Washington University in St. Louis,* [f]*lei.liu@wustl.edu*

[6]*School of Mathematics, Shandong University,* [g]*yinyue0607@163.com*

[7]*Beijing International Center for Mathematical Research and Department of Biostatistics and Chongqing Research Institute of Big Data, Peking University,* [h]*azhou@math.pku.edu.cn*

In this study we focus on Cooperative Differential Network Learning with hub detection (CDNL) for functional Magnetic Resonance Imaging (fMRI) scan from multiple research centers. As research centers may use varying scanners, imaging parameters, and other conditions that introduce heterogeneity, CDNL allows us to analyze fMRI data from various perspectives while identifying shared structures, potentially revealing the underlying mechanisms of neurological diseases. In addition, brain functional networks often consist of multiple hubs-central nodes within the network that play a crucial role in supporting integrated brain function. Investigating these hubs can offer valuable insight into functional connectivity patterns in the brain. To address this task, we formulate it as a penalized logistic regression problem and introduce two independent penalties (Cooperative Penalty and Hub Penalty) to enable simultaneous estimation of multiple differential networks with hub detection. To further enhance empirical performance, we develop an ensemble-learning procedure. We conduct comprehensive simulation studies to assess the finite-sample performance of the proposed method and compare it with existing state-of-the-art alternatives. In the application we apply the proposed method to analyze multiple fMRI scans related to Attention Deficit Hyperactivity Disorder from various research centers. We identify common hub brain regions and similar differential interaction patterns across various centers. These findings are highly consistent with existing results from clinical medical research.

REFERENCES

ABROL, A., DAMARAJU, E., MILLER, R. L., STEPHEN, J. M., CLAUS, E. D., MAYER, A. R. and CALHOUN, V. D. (2017). Replicability of time-varying connectivity patterns in large resting state fMRI samples. *NeuroImage* **163** 160–176. https://doi.org/10.1016/j.neuroimage.2017.09.020

BELLEC, P., CHU, C., CHOUINARD-DECORTE, F., BENHAJALI, Y., MARGULIES, D. S. and CRADDOCK, R. C. (2017). The neuro bureau ADHD-200 preprocessed repository. *NeuroImage* **144** 275–286.

BOS, D. J., ORANJE, B., ACHTERBERG, M., VLASKAMP, C., AMBROSINO, S., DE REUS, M. A., VAN DEN HEUVEL, M. P., ROMBOUTS, S. A. R. B. and DURSTON, S. (2017). Structural and functional connectivity in children and adolescents with and without attention deficit/hyperactivity disorder. *J. Child Psychol. Psychiatry* **58** 810–818. https://doi.org/10.1111/jcpp.12712

BUTTON, K. S., IOANNIDIS, J. P., MOKRYSZ, C., NOSEK, B. A., FLINT, J., ROBINSON, E. S. and MUNAFÒ, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14** 365–376.

Key words and phrases. Brain functional connectivity, Cooperative Learning with Hub Detection, Differential Network, heterogeneous fMRI data, network comparison.

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973 https://doi.org/10.1198/jasa.2011.tm10155

CHEN, H., GUO, D., GUO, Y., HE, Y., LIU, D., LIU, L., YIN, Y. and ZHOU, X.-H. (2025). Supplement to "Cooperative Differential Network Learning with Hub Detection for Multicenter Neuroimaging data." https://doi.org/10.1214/25-AOAS2026SUPPA, https://doi.org/10.1214/25-AOAS2026SUPPB

CHEN, H., GUO, Y., HE, Y., JI, J., LIU, L., SHI, Y., WANG, Y., YU, L., ZHANG, X. et al. (2022). Simultaneous differential network analysis and classification for matrix-variate data with application to brain connectivity. *Biostatistics* **23** 967–989. MR4454965 https://doi.org/10.1093/biostatistics/kxab007

COHEN, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Routledge, London.

DICKERSON, B. C. (2007). Advances in functional magnetic resonance imaging: Technology and clinical applications. *Neurotherapeutics* **4** 360–370. https://doi.org/10.1016/j.nurt.2007.05.007

FORNITO, A., ZALESKY, A. and BREAKSPEAR, M. (2015). The connectomics of brain disorders. *Nat. Rev. Neurosci.* **16** 159–172. https://doi.org/10.1038/nrn3901

GRADY, C. L., RIECK, J. R., NICHOL, D., RODRIGUE, K. M. and KENNEDY, K. M. (2021). Influence of sample size and analytic approach on stability and interpretation of brain-behavior correlations in task-related fMRI data. *Hum. Brain Mapp.* **42** 204–219. https://doi.org/10.1002/hbm.25217

GUO, X., YAO, D., CAO, Q., LIU, L., ZHAO, Q., LI, H., HUANG, F., WANG, Y., QIAN, Q. et al. (2020). Shared and distinct resting functional connectivity in children and adults with attention-deficit/hyperactivity disorder. *Transl. Psychiatry* **10** 65.

HE, Y., JI, J., XIE, L., ZHANG, X. and XUE, F. (2018). A new insight into underlying disease mechanism through semi-parametric latent differential network model. *BMC Bioinform.* **19** 5–19.

HORN, J. D. V. and TOGA, A. W. (2009). Multisite neuroimaging trials. *Curr. Opin. Neurol.* **22** 370–378. https://doi.org/10.1097/WCO.0b013e32832d92de

HU, W., SHEN, W., ZHOU, H. and KONG, D. (2020). Matrix linear discriminant analysis. *Technometrics* **62** 196–205. MR4095745 https://doi.org/10.1080/00401706.2019.1610069

JI, J., HE, D., FENG, Y., HE, Y., XUE, F. and XIE, L. (2017). JDINAC: Joint density-based non-parametric differential interaction network analysis and classification using high-dimensional sparse omics data. *Bioinformatics* **33** 3080–3087.

JI, J., HE, Y., LIU, L. and XIE, L. (2021). Brain connectivity alteration detection via matrix-variate differential network model. *Biometrics* **77** 1409–1421. MR4357847 https://doi.org/10.1111/biom.13359

MURIAS, M., SWANSON, J. M. and SRINIVASAN, R. (2007). Functional connectivity of frontal cortex in healthy and ADHD children reflected in EEG coherence. *Cereb. Cortex* **17** 1788–1799.

NARAYAN, M. and ALLEN, G. I. (2016). Mixed effects models for resampled network statistics improves statistical power to find differences in multi-subject functional connectivity. *Front. Neurosci.* **10** 108. https://doi.org/10.3389/fnins.2016.00108

OU-YANG, L., ZHANG, X.-F., ZHAO, X.-M., WANG, D. D., WANG, F. L., LEI, B. and YAN, H. (2018). Joint learning of multiple differential networks with latent variables. *IEEE Trans. Cybern.* **49** 3494–3506.

QIU, Y. and ZHOU, X.-H. (2020). Estimating c-level partial correlation graphs with application to brain imaging. *Biostatistics* **21** 641–658. MR4164049 https://doi.org/10.1093/biostatistics/kxy076

QIU, Y. and ZHOU, X.-H. (2022). Inference on multi-level partial correlations based on multi-subject time series data. *J. Amer. Statist. Assoc.* **117** 2268–2282. MR4528503 https://doi.org/10.1080/01621459.2021.1917417

SAWILOWSKY, S. S. (2009). New effect size rules of thumb. *J. Mod. Appl. Stat. Methods* **8** 597–599.

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

SMITH, S. M. (2012). The future of FMRI connectivity. *NeuroImage* **62** 1257–1266. https://doi.org/10.1016/j.neuroimage.2012.01.022

STOODLEY, C. J. (2016). The cerebellum and neurodevelopmental disorders. *Cerebellum* **15** 34–37. https://doi.org/10.1007/s12311-015-0715-3

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

THE ADHD-200 DATASET WEBSITE (2019). The Neuro Bureau. Available at http://neurobureau.projects.nitrc.org/ADHD200/Data.html (Accessed March 7, 2019).

VAIDYA, C. J. (2012). Neurodevelopmental abnormalities in ADHD. In *Behavioral Neuroscience of Attention Deficit Hyperactivity Disorder and Its Treatment* 49–66.

VALERA, E. M., FARAONE, S. V., MURRAY, K. E. and SEIDMAN, L. J. (2007). Meta-analysis of structural imaging findings in attention-deficit/hyperactivity disorder. *Biol. Psychiatry* **61** 1361–1369. https://doi.org/10.1016/j.biopsych.2006.06.011

VAN DEN HEUVEL, M. P. and SPORNS, O. (2013). Network hubs in the human brain. *Trends Cogn. Sci.* **17** 683–696. https://doi.org/10.1016/j.tics.2013.09.012

WANG, B., WANG, G., WANG, X., CAO, R., XIANG, J., YAN, T., LI, H., YOSHIMURA, S., TOICHI, M. et al. (2021). Rich-club analysis in adults with ADHD connectomes reveals an abnormal structural core network. *J. Atten. Disord.* **25** 1068–1079.

WANG, Y., KANG, J., KEMMER, P. B. and GUO, Y. (2016). An efficient and reliable statistical method for estimating functional connectivity in large scale brain networks using partial correlation. *Front. Neurosci.* **10** 179959.

XIA, M., WANG, J. and HE, Y. (2013). BrainNet Viewer: A network visualization tool for human brain connectomics. *PLoS ONE* **8** e68910.

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

YU, L., HE, Y., KONG, X. and ZHANG, X. (2022). Projected estimation for large-dimensional matrix factor models. *J. Econometrics* **229** 201–217. MR4414019 https://doi.org/10.1016/j.jeconom.2021.04.001

ZHAN, K., ZHANG, C., GUAN, J. and WANG, J. (2017). Graph learning for multiview clustering. *IEEE Trans. Cybern.* **48** 2887–2895.

ZHANG, X.-F., OU-YANG, L. and YAN, H. (2017). Incorporating prior information into differential network analysis using non-paranormal graphical models. *Bioinformatics* **33** 2436–2445. https://doi.org/10.1093/bioinformatics/btx208

ZHU, Y. and LI, L. (2018). Multiple matrix Gaussian graphs estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 927–950. MR3874304 https://doi.org/10.1111/rssb.12278

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1603–1622
https://doi.org/10.1214/24-AOAS1993

A MULTIRUN STEP-STRESS MODEL FOR TREND RENEWAL DATA WITH APPLICATIONS TO LIFETIME ASSESSMENT FOR RECHARGEABLE BATTERIES

BY SHENG-TSAING TSENG[1,a], NAN-JUNG HSU[1,b] AND CHIEN-CHI WU[1,c]

[1]*Institute of Statistics and Data Science, National Tsing-Hua University,* [a]*sttseng@stat.nthu.edu.tw,* [b]*njhsu@stat.nthu.edu.tw,* [c]*nicknick0826@gmail.com*

Conducting a cost-efficient lifetime-testing plan to timely assess lifetime information of a highly reliable product is often a challenging task in the manufacturing industry. Motivated by a case study of rechargeable lithium-ion batteries, this paper introduces a multirun k-level step-stress experiment, running under different stresses in repeated cycles, to collect and analyze the degradation data of reusable highly-reliable products. Specifically, we formulate the battery capacity over recharge cycles as a counting process and adopt a trend renewal process (TRP) to characterize the degradation patterns of capacity varying with the stress level of the accelerated factor. By using a Markovian property on cumulative exposure in our counting process, the degradation data observed in a multirun k-level step-stress TRP model can be converted equivalently to corresponding k constant-stress TRP models. This connection allows us to estimate the parameters using maximum likelihood and to infer with uncertainty quantification the end-of-performance (EOP) of batteries at normal-use conditions. This novel method is shown to be efficient for the lifetime assessment of reusable products.

REFERENCES

CUI, D., SUN, Q. and XIE, M. (2023). Robust statistical modeling of heterogeneity for repairable systems using multivariate Gaussian convolution processes. *IEEE Trans. Reliab.* **72** 1493–1506.

DICICCIO, T. J. and EFRON, B. (1996). Bootstrap confidence intervals. *Statist. Sci.* **11** 189–228. With comments and a rejoinder by the authors. MR1436647 https://doi.org/10.1214/ss/1032280214

HU, C.-H., LEE, M.-Y. and TANG, J. (2015). Optimum step-stress accelerated degradation test for Wiener degradation process under constraints. *European J. Oper. Res.* **241** 412–421. MR3276771 https://doi.org/10.1016/j.ejor.2014.09.003

LEE, I., HONG, Y., TSENG, S.-T. and DASGUPTA, T. (2018). Sequential Bayesian design for accelerated life tests. *Technometrics* **60** 472–483. MR3878102 https://doi.org/10.1080/00401706.2018.1437475

LEE, I., TSENG, S.-T. and HONG, Y. (2020). Global planning of accelerated degradation tests based on exponential dispersion degradation models. *Naval Res. Logist.* **67** 469–483. MR4157266 https://doi.org/10.1002/nav.21923

LI, Y., LIU, K., FOLEY, A. M., ZLKE, A., BERECIBAR, M., NANINE-MAURY, E., VAN MIERLO, J. and HOSTER, H. E. (2019). Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review. *Renew. Sustain. Energy Rev.* **113** 109254. Elsevier.

LIAO, C. M. and TSENG, S. T. (2006). Optimal design for step-stress accelerated degradation tests. *IEEE Trans. Reliab.* **55** 59–66.

LINDQVIST, B. H., ELVEBAKK, G. and HEGGLAND, K. (2003). The trend-renewal process for statistical analysis of repairable systems. *Technometrics* **45** 31–44. MR1956189 https://doi.org/10.1198/004017002188618671

MEEKER, W. Q. and ESCOBAR, L. A. (1998). *Statistical Methods for Reliability Data*. Wiley, New York.

MIN, J., HONG, Y., KING, C. B. and MEEKER, W. Q. (2022). Reliability analysis of artificial intelligence systems using recurrent events data from autonomous vehicles. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **71** 987–1013. MR4470829 https://doi.org/10.1111/rssc.12564

NELSON, W. (1990). *Accelerated Testing: Statistical Models, Testing Plan and Data Analysis*. Wiley, New York.

NG, S. S. Y., 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.

Key words and phrases. Trend renewal process, trend renewal data, a multirun step-stress experiment, lithium-ion batteries, end of performance.

PARK, J. I. and BAE, S. J. (2010). Direct prediction methods on lifetime distribution of organic light-emitting diodes from accelerated degradation tests. *IEEE Trans*. *Reliab*. **59** 74–90.

PENG, C. Y. and TSENG, S. T. (2010). Progressive-stress accelerated degradation test for highly-reliable products. *IEEE Trans*. *Reliab*. **59** 30–37.

RIGDON, S. E. and BASU, A. P. (2000). *Statistical Methods for the Reliability of Repairable Systems*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. Wiley, New York. A Wiley-Interscience Publication. MR1781523

ROSS, S. M. (2019). *Introduction to Probability Models*. Academic Press, London. Twelfth edition of [MR0328973]. MR3931305

SHAN, Q., HONG, Y. and MEEKER, W. Q. (2020). Seasonal warranty prediction based on recurrent event data. *Ann*. *Appl*. *Stat*. **14** 929–955. MR4117835 https://doi.org/10.1214/20-AOAS1333

TSENG, S.-T. and LEE, I. (2016). Optimum allocation rule for accelerated degradation tests with a class of exponential-dispersion degradation models. *Technometrics* **58** 244–254. MR3488303 https://doi.org/10.1080/00401706.2015.1033109

TSENG, S. T., BALAKRISHNAN, N. and TSAI, C. C. (2009). Optimal step-stress accelerated degradation test plan for gamma degradation processes. *IEEE Trans*. *Reliab*. **58** 611–618.

TSENG, S.-T., HSU, N.-J. and WU, C.-C. (2025). Supplement to "A MultiRun Step-Stress Model for Trend Renewal Data with Applications to Lifetime Assessment for Rechargeable Batteries." https://doi.org/10.1214/24-AOAS1993SUPPA, https://doi.org/10.1214/24-AOAS1993SUPPB

TSENG, S. T. and WEN, Z. C. (2000). Step-stress accelerated degradation analysis for highly reliable products. *J*. *Qual*. *Technol*. **32** 209–216.

TUNG, H.-P. and TSENG, S.-T. (2019). Planning gamma accelerated degradation tests with two accelerating variables. *Naval Res*. *Logist*. **66** 439–447. MR3979802 https://doi.org/10.1002/nav.21848

WANG, Y. F., TSENG, S. T., LINDQUVIST, B. H. and TSUI, K. L. (2019). End of performance prediction of lithium-ion batteries. *J*. *Qual*. *Technol*. **51** 198–213.

XIAO, S., KOTTAS, A., SANSÓ, B. and KIM, H. (2021). Nonparametric Bayesian modeling and estimation for renewal processes. *Technometrics* **63** 100–115. MR4205695 https://doi.org/10.1080/00401706.2019.1693428

YE, Z. S., CHEN, L. P., TANG, L. C. and XIE, M. (2014). Accelerated degradation test planning using the inverse Gaussian process. *IEEE Trans*. *Reliab*. **63** 750–763.

ZHANG, J. L. and LEE, J. (2011). A review on prognostics and health monitoring of Li-ion battery. *J*. *Power Sources* **196** 6007–6014.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1623–1640
https://doi.org/10.1214/24-AOAS1997

AN APPLICATION OF VINE-BASED REGRESSION TO FLIGHT LANDING DATA

BY HASSAN ALNASSER[a] AND CLAUDIA CZADO[b]

Applied Mathematical Statistics, Technische Universität München, [a]*h.alnasser@tum.de,* [b]*cczado@ma.tum.de*

Runway overruns pose a significant challenge to aviation safety and represent the most common type of landing incident. Identifying the underlying factors contributing to runway overruns is vital for the development of effective prevention strategies. Since a few overruns are observed, the distance required for a pilot to achieve control of the aircraft, measured by the distance to decelerate to 80 knots, is used as a precursor. We propose vine copula-based distributional regression models for the distance to 80 knots, given a set of contributing factors, for the analysis of 711 flights sourced from the quick access recorder (QAR). This approach allows to accommodate non-linear, non-Gaussian patterns observed in the data and to rank the impact of the contributing factors. Three vine-based regression models and two Gaussian benchmark models were investigated, revealing that D-vine regression model is the preferred model. This D-vine regression is then used to identify 41 high-risk flights, using estimated conditional probabilities for the distance to controllable speed exceeding a significant threshold. The analysis of the contributing factors for these flights reveals distinct marginal behaviors and dependent patterns that must be considered when designing risk prevention strategies.

REFERENCES

AAS, K. (2016). Pair-copula constructions for financial applications: A review. *Econometrics* **4** 43.

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

ALNASSER, H. and CZADO, C. (2025). Supplement to "An application of vine-based regression to flight landing data." https://doi.org/10.1214/24-AOAS1997SUPP

ANDERSON, J. D. (2007). *Fundamentals of Aerodynamics*. *McGraw-Hill Series in Aeronautical and McGraw-Hill Higher Education*.

ARNALDO VALDÉS, R. M., GÓMEZ COMENDADOR, V. F., PEREZ SANZ, L. and RODRIGUEZ SANZ, A. (2018). Prediction of aircraft safety incidents using Bayesian inference and hierarchical structures. *Saf. Sci.* **104** 216–230. https://doi.org/10.1016/j.ssci.2018.01.008

AU, S.-K. and BECK, J. L. (2001). Estimation of small failure probabilities in high dimensions by subset simulation. *Probab. Eng. Mech.* **16** 263–277. https://doi.org/10.1016/S0266-8920(01)00019-4

AYRA, E. S., RÍOS INSUA, D. and CANO, J. (2019). Bayesian network for managing runway overruns in aviation safety. *J. Aerosp. Inform. Syst.* **16** 546–558.

BEDFORD, T. J. and COOKE, R. (2001). Probability density decomposition for conditionally dependent random variables modeled by Vines. *Annals of Mathematics and Artificial Intigence* **32** 245–268. MR1859866 https://doi.org/10.1023/A:1016725902970

BURIN, J. M. (2011a). Keys to a safe arrival. https://flightsafety.org/asw-article/keys-to-a-safe-arrival/, Last accessed on 2022-02-25.

BURIN, J. M. (2011b). Keys to a safe arrival. *AeroSafety World* **6**.

CHU, L., LIU, Y. and FWA, T. (2024). Estimating runway overrun risks of landing aircraft. *J. Infrastruct. Syst.* **30** 04024002.

COX, L. A. (2008). What's wrong with risk matrices? *Risk Anal.* **28** 497–512.

CZADO, C. (2010). Pair-copula constructions of multivariate copulas. In *Copula Theory and Its Applications*. *Lect. Notes Stat. Proc.* **198** 93–109. Springer, Heidelberg. MR3051264 https://doi.org/10.1007/978-3-642-12465-5_4

Key words and phrases. Conditional probability estimation, D-vine regression, non-Gaussian dependencies, runway overruns, uncertainty quantification, risky flights.

CZADO, C. (2019). *Analyzing Dependent Data with Vine Copulas*: *A Practical Guide with R*. *Lecture Notes in Statistics* **222**. Springer, Cham. MR3931334 https://doi.org/10.1007/978-3-030-13785-4

CZADO, C. and NAGLER, T. (2022). Vine copula based modeling. *Annu*. *Rev*. *Stat*. *Appl*. **9** 453–477. MR4394916 https://doi.org/10.1146/annurev-statistics-040220-101153

DISSMANN, J., BRECHMANN, E. C., CZADO, C. and KUROWICKA, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Comput*. *Statist*. *Data Anal*. **59** 52–69. MR3000041 https://doi.org/10.1016/j.csda.2012.08.010

DREES, L. (2016). Predictive analysis: Quantifying operational airline risks. Dissertation, Technische Univ. München, München.

DREES, L., WANG, C. and HOLZAPFEL, F. (2014). Using subset simulation to quantify stakeholder contribution to runway overrun. In *Proceedings of Probabilistic Safety Assessment and Management PSAM* **12**.

GU, R. and WANG, P. (2014). Estimation of wet and contaminated runway landing distance based on multiple linear regression. *Journal of Civil Aviation University of China* **32** 20.

HERRMANN, J. S. (2018). Regular vine copula based quantile regression. Master Thesis in Mathematics, Technical Univ. Munich, Munich.

HU, C., ZHOU, S.-H., XIE, Y. and CHANG, W.-B. (2016). The study on hard landing prediction model with optimized parameter SVM method. In 2016 35*th Chinese Control Conference* (*CCC*) 4283–4287. IEEE.

IATA (2022). Air passenger numbers to recover in 2024. https://www.iata.org/en/pressroom/2022-releases/2022-03-01-01/, Last accessed on 2022-03-09.

ICAO (2006). ICAO annex 19, safety management. https://www.icao.int/safety/SafetyManagement/, Last accessed on 2022-02-25.

ICAO (2013). *Safety Management Manual* (*SMS*), 3rd edn. Doc 9859.

JOE, H. (2014). *Dependence Modeling with Copulas*. *Monographs on Statistics and Applied Probability* **134**. CRC Press, New York. https://doi.org/10.1201/b17116

JOE, H. and XU, J. J. (1996). The estimation method of inference functions for margins for multivariate models. https://doi.org/10.14288/1.0225985

KLEIN, N. (2024). Distributional regression for data analysis. *Annu*. *Rev*. *Stat*. *Appl*. **11** 321–346. MR4742287

KRAUS, D. and CZADO, C. (2017). D-vine copula based quantile regression. *Comput*. *Statist*. *Data Anal*. **110** 1–18. MR3612604 https://doi.org/10.1016/j.csda.2016.12.009

NAGLER, T. (2021). Vinereg: D-vine quantile regression. R package version 0.7.4.

NAGLER, T. and VATTER, T. (2021). Rvinecopulib: High performance algorithms for vine copula modeling. R package version 0.5.5.1.1.

PATTON, A. J. (2012). A review of copula models for economic time series. *J*. *Multivariate Anal*. **110** 4–18. MR2927506 https://doi.org/10.1016/j.jmva.2012.02.021

SCHOLZ, D. (2012). *Aircraft Design*. Springer.

SFORZA, P. M. (2014). *Commercial Airplane Design Principles*. *Elsevier Aerospace Engineering Series*. Elsevier Science.

SKLAR, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ*. *Inst*. *Stat*. *Univ*. *Paris* **8** 229–231. MR0125600

STÖBER, J., JOE, H. and CZADO, C. (2013). Simplified pair copula constructions—limitations and extensions. *J*. *Multivariate Anal*. **119** 101–118. MR3061418 https://doi.org/10.1016/j.jmva.2013.04.014

STRAUB, D. and DER KIUREGHIAN, A. (2010). Bayesian network enhanced with structural reliability methods: Application. *J*. *Eng*. *Mech*. **136** 1259–1270.

TANG, X.-S., LI, D.-Q., ZHOU, C.-B. and PHOON, K.-K. (2015). Copula-based approaches for evaluating slope reliability under incomplete probability information. *Struct*. *Saf*. **52** 90–99.

TEPEGJOZOVA, M., ZHOU, J., CLAESKENS, G. and CZADO, C. (2022). Nonparametric C- and D-vine-based quantile regression. *Depend*. *Model*. **10** 1–21. MR4386473 https://doi.org/10.1515/demo-2022-0100

VALDÉS, R. M. A., COMENDADOR, F. G., GORDÚN, L. M. and NIETO, F. J. S. (2011). The development of probabilistic models to estimate accident risk (due to runway overrun and landing undershoot) applicable to the design and construction of runway safety areas. *Saf*. *Sci*. **49** 633–650.

WAGNER, D. C. and BARKER, K. (2014). Statistical methods for modeling the risk of runway excursions. *J*. *Risk Res*. **17** 885–901.

WANG, L., WU, C. and SUN, R. (2014). An analysis of flight quick access recorder (QAR) data and its applications in preventing landing incidents. *Reliab*. *Eng*. *Syst*. *Saf*. **127** 86–96.

WANG, X., FANG, X., BELLER, L. and HOLZAPFEL, F. (2020). Calibration of contributing factors for model-based predictive analysis algorithm using polynomial chaos expansion methods. https://doi.org/10.3850/978-981-14-8593-0_3740-cd

WONG, D. K., PITFIELD, D., CAVES, R. E. and APPLEYARD, A. (2006). Quantifying and characterising aviation accident risk factors. *J*. *Air Transp*. *Manag*. **12** 352–357.

YOU, X., JI, M. and HAN, H. (2013). The effects of risk perception and flight experience on airline pilots' locus of control with regard to safety operation behaviors. *Accident Anal. Prev.* **57** 131–139.

ZHAO, N. and ZHANG, J. (2022). Research on the prediction of aircraft landing distance. *Math. Probl. Eng.* **2022**.

ZWIRGLMAIER, K. and STRAUB, D. (2016). A discretization procedure for rare events in Bayesian networks. *Reliab. Eng. Syst. Saf.* **153** 96–109.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1641–1662
https://doi.org/10.1214/24-AOAS1998

LEARN THEN TEST: CALIBRATING PREDICTIVE ALGORITHMS TO ACHIEVE RISK CONTROL

BY ANASTASIOS N. ANGELOPOULOS[1,a], STEPHEN BATES[2,c], EMMANUEL J. CANDÈS[3,d], MICHAEL I. JORDAN[1,b] AND LIHUA LEI[4,e]

[1]*Department of Electrical Engineering and Computer Science, University of California, Berkeley,* [a]*angelopoulos@berkeley.edu,* [b]*jordan@cs.berkeley.edu*

[2]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology,* [c]*stephenbates@mit.edu*

[3]*Department of Statistics, Stanford University,* [d]*candes@stanford.edu*

[4]*Stanford Graduate School of Business,* [e]*lihualei@stanford.edu*

We introduce a framework for calibrating machine learning models to satisfy finite-sample statistical guarantees. Our calibration algorithms work with any model and (unknown) data-generating distribution and do not require retraining. The algorithms address, among other examples, false discovery rate control in multilabel classification, intersection-over-union control in instance segmentation, and simultaneous control of the type-1 outlier error and confidence set coverage in classification or regression. Our main insight is to reframe risk control as multiple hypothesis testing, enabling different mathematical arguments. We demonstrate our algorithms with detailed worked examples in computer vision and tabular medical data. The computer vision experiments demonstrate the utility of our approach in calibrating state-of-the-art predictive architectures that have been deployed widely, such as the `detectron2` object detection system.

REFERENCES

ANGELOPOULOS, A. N., BATES, S., CANDÈS, E. J., JORDAN, M. I. and LEI, L. (2025). Supplement to "Learn then test: Calibrating predictive algorithms to achieve risk control." https://doi.org/10.1214/24-AOAS1998SUPPA, https://doi.org/10.1214/24-AOAS1998SUPPB

ANGELOPOULOS, A. N., BATES, S., MALIK, J. and JORDAN, M. I. (2021). Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations* (*ICLR*).

ANGELOPOULOS, A. N., BATES, S. et al. (2023). Conformal prediction: A gentle introduction. *Found*. *Trends Mach*. *Learn*. **16** 494–591.

ANGELOPOULOS, A. N., BATES, S., ZRNIC, T. and JORDAN, M. I. (2022). Private prediction sets. *Harv. Data Sci*. *Rev*. **4**.

ANTHONY, M. and SHAWE-TAYLOR, J. (1993). A result of Vapnik with applications. *Discrete Appl*. *Math*. **47** 207–217. MR1255199 https://doi.org/10.1016/0166-218X(93)90126-9

BARBER, R. F., CANDÈS, E. J., RAMDAS, A. and TIBSHIRANI, R. J. (2021a). Predictive inference with the jackknife+. *Ann*. *Statist*. **49** 486–507. MR4206687 https://doi.org/10.1214/20-AOS1965

BARBER, R. F., CANDÈS, E. J., RAMDAS, A. and TIBSHIRANI, R. J. (2021b). The limits of distribution-free conditional predictive inference. *Inf. Inference* **10** 455–482. MR4270755 https://doi.org/10.1093/imaiai/iaaa017

BATES, S., ANGELOPOULOS, A., LEI, L., MALIK, J. and JORDAN, M. (2021). Distribution-free, risk-controlling prediction sets. *J. ACM* **68** Art. 43. MR4402354 https://doi.org/10.1145/3478535

BATES, S., CANDÈS, E., LEI, L., ROMANO, Y. and SESIA, M. (2023). Testing for outliers with conformal p-values. *Ann*. *Statist*. **51** 149–178. MR4564852 https://doi.org/10.1214/22-aos2244

BAUER, P. (1991). Multiple testing in clinical trials. *Stat*. *Med*. **10** 871–890.

BENTKUS, V. (2004). On Hoeffding's inequalities. *Ann*. *Probab.* **32** 1650–1673. MR2060313 https://doi.org/10.1214/009117904000000360

BREIMAN, L. (2001). Statistical modeling: The two cultures. *Statist*. *Sci*. **16** 199–231. MR1874152 https://doi.org/10.1214/ss/1009213726

BRETZ, F., MAURER, W., BRANNATH, W. and POSCH, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Stat*. *Med*. **28** 586–604. MR2655732 https://doi.org/10.1002/sim.3495

Key words and phrases. Conformal prediction, deep learning, computer vision, machine learning.

CANDÈS, E., LEI, L. and REN, Z. (2023). Conformalized survival analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 24–45. MR4716146 https://doi.org/10.1093/jrsssb/qkac004

CAUCHOIS, M., GUPTA, S., ALI, A. and DUCHI, J. C. (2024). Robust validation: Confident predictions even when distributions shift. *J. Amer. Statist. Assoc.* **119** 3033–3044. MR4833935 https://doi.org/10.1080/01621459.2023.2298037

CAUCHOIS, M., GUPTA, S. and DUCHI, J. C. (2021). Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. *J. Mach. Learn. Res.* **22** Paper No. 81. MR4253774

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448 https://doi.org/10.1214/13-AOS1161

CHERNOZHUKOV, V., WÜTHRICH, K. and YINCHU, Z. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Proceedings of the* 31*st Conference on Learning Theory*. *PMLR* **75** 732–749.

COHEN, J. W., COHEN, S. B. and BANTHIN, J. S. (2009). The medical expenditure panel survey: A national information resource to support healthcare cost research and inform policy and practice. *Med. Care* S44–S50.

DENG, J. et al. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition* 248–255.

DUNN, R., WASSERMAN, L. and RAMDAS, A. (2020). Distribution-free prediction sets with random effects. ArXiv preprint. Available at arXiv:1809.07441.

EFRON, B. (2020). Prediction, estimation, and attribution. *Int. Stat. Rev.* **88** S28–S59.

FANG, Z. and SANTOS, A. (2019). Inference on directionally differentiable functions. *Rev. Econ. Stud.* **86** 377–412. MR3936869 https://doi.org/10.1093/restud/rdy049

FISCH, A., SCHUSTER, T., JAAKKOLA, T. and BARZILAY, R. (2021a). Few-shot conformal prediction with auxiliary tasks. ArXiv preprint. Available at arXiv:2102.08898.

FISCH, A., SCHUSTER, T., JAAKKOLA, T. S. and BARZILAY, R. (2021b). Efficient conformal prediction via cascaded inference with expanded admission. In *International Conference on Learning Representations*.

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451

GIBBS, I. and CANDÈS, E. (2021). Adaptive conformal inference under distribution shift. ArXiv preprint. Available at arXiv:2106.00170.

GUAN, L. (2019). Conformal prediction with localization. ArXiv preprint. Available at arXiv:1908.08558.

GUAN, L. and TIBSHIRANI, R. (2022). Prediction and outlier detection in classification problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 524–546. MR4412996 https://doi.org/10.1111/rssb.12443

HE, K., GKIOXARI, G., DOLLÁR, P. and GIRSHICK, R. (2017). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision* 2961–2969.

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

HECHTLINGER, Y., PÓCZOS, B. and WASSERMAN, L. (2018). Cautious deep learning. ArXiv preprint. Available at arXiv:1805.09460.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

HU, X. and LEI, J. (2020). A distribution-free test of covariate shift using conformal prediction. ArXiv preprint. Available at arXiv:2010.07147.

IZBICKI, R., SHIMIZU, G. and STERN, R. (2020). Flexible distribution-free conditional predictive bands using density estimators. In *International Conference on Artificial Intelligence and Statistics* 3068–3077. PMLR.

KRISHNAMOORTHY, K. and MATHEW, T. (2009). *Statistical Tolerance Regions*: *Theory*, *Applications*, *and Computation*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2500599 https://doi.org/10.1002/9780470473900

KRIZHEVSKY, A., HINTON, G. et al. (2009). Learning multiple layers of features from tiny images.

KUCHIBHOTLA, A. K. and BERK, R. A. (2023). Nested conformal prediction sets for classification with applications to probation data. *Ann. Appl. Stat.* **17** 761–785. MR4539052 https://doi.org/10.1214/22-aoas1650

LEI, J. (2014). Classification with confidence. *Biometrika* **101** 755–769. MR3286915 https://doi.org/10.1093/biomet/asu038

LEI, J., RINALDO, A. and WASSERMAN, L. (2015). A conformal prediction approach to explore functional data. *Ann. Math. Artif. Intell.* **74** 29–43. MR3353895 https://doi.org/10.1007/s10472-013-9366-6

LEI, L. and CANDÈS, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 911–938. MR4349122

LIANG, S., LI, Y. and SRIKANT, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. ArXiv preprint. Available at arXiv:1706.02690.

LIN, T.-Y., GOYAL, P., GIRSHICK, R., HE, K. and DOLLÁR, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision* 2980–2988.

LIN, T.-Y., MAIRE, M., BELONGIE, S., HAYS, J., PERONA, P., RAMANAN, D., DOLLÁR, P. and ZITNICK, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision* 740–755. Springer, Berlin.

NGUYEN, D. T., PATHAK, R., ANGELOPOULOS, A. N., BATES, S. and JORDAN, M. I. (2024). Data-Adaptive Tradeoffs among Multiple Risks in Distribution-Free Prediction. ArXiv preprint. Available at arXiv:2403.19605.

PAPADOPOULOS, H., PROEDROU, K., VOVK, V. and GAMMERMAN, A. (2002). Inductive confidence machines for regression. In *Machine Learning*: *ECML* 2002. *Lecture Notes in Computer Science* **2430** 345–356. Springer, Berlin. MR2050303 https://doi.org/10.1007/3-540-36755-1_29

PARK, S., BASTANI, O., MATNI, N. and LEE, I. (2020). PAC confidence sets for deep neural networks via calibrated prediction. In *ICLR* 2020.

PARK, S., DOBRIBAN, E., LEE, I. and BASTANI, O. (2021). PAC prediction sets under covariate shift. ArXiv preprint. Available at arXiv:2106.09848.

ROMANO, Y., PATTERSON, E. and CANDES, E. (2019). Conformalized quantile regression. *Adv. Neural Inf. Process. Syst.* **32**.

ROMANO, Y., SESIA, M. and CANDÈS, E. (2020). Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems* **33** 3581–3591.

SADINLE, M., LEI, J. and WASSERMAN, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *J. Amer. Statist. Assoc.* **114** 223–234. MR3941250 https://doi.org/10.1080/01621459.2017.1395341

SCHUSTER, T., FISCH, A., JAAKKOLA, T. and BARZILAY, R. (2021). Consistent accelerated inference via confident adaptive transformers. ArXiv preprint. Available at arXiv:2104.08803.

TIBSHIRANI, R. J., FOYGEL BARBER, R., CANDES, E. and RAMDAS, A. (2019). Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems* **32** 2530–2540.

TUKEY, J. W. (1947). Non-parametric estimation. II. Statistically equivalent blocks and tolerance regions–the continuous case. *Ann. Math. Stat.* **18** 529–539. MR0023033 https://doi.org/10.1214/aoms/1177730343

VIOLA, P. and JONES, M. J. (2004). Robust real-time face detection. *Int. J. Comput. Vis.* **57** 137–154.

VOVK, V. (2015). Cross-conformal predictors. *Ann. Math. Artif. Intell.* **74** 9–28. MR3353894 https://doi.org/10.1007/s10472-013-9368-4

VOVK, V. (2019). Universally consistent conformal predictive distributions. In *Proceedings of the Eighth Symposium on Conformal and Probabilistic Prediction and Applications*. *PMLR* **105** 105–122.

VOVK, V. (2021). Testing randomness online. *Statist. Sci.* **36** 595–611. MR4323055 https://doi.org/10.1214/20-sts817

VOVK, V. and BENDTSEN, C. (2018). Conformal predictive decision making. In *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*. *Proceedings of Machine Learning Research* **91** 52–62.

VOVK, V., GAMMERMAN, A. and SAUNDERS, C. (1999). Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning* 444–453.

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York. MR2161220

VOVK, V., PETEJ, I., TOCCACELI, P., GAMMERMAN, A., AHLBERG, E. and CARLSSON, L. (2020). Conformal calibrators. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications* **128** 84–99.

WALD, A. (1943). An extension of Wilks' method for setting tolerance limits. *Ann. Math. Stat.* **14** 45–55. MR0007965 https://doi.org/10.1214/aoms/1177731491

WAUDBY-SMITH, I. and RAMDAS, A. (2020). Variance-adaptive confidence sequences by betting. ArXiv preprint. Available at arXiv:2010.09686.

WILKS, S. S. (1941). Determination of sample sizes for setting tolerance limits. *Ann. Math. Stat.* **12** 91–96. MR0004451 https://doi.org/10.1214/aoms/1177731788

WILKS, S. S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *Ann. Math. Stat.* **13** 400–409. MR0007592 https://doi.org/10.1214/aoms/1177731537

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1663–1690
https://doi.org/10.1214/24-AOAS2002

BAYESIAN INFERENCE FOR PARTIAL ORDERS FROM RANDOM LINEAR EXTENSIONS: POWER RELATIONS FROM 12TH CENTURY ROYAL ACTA

BY GEOFF K. NICHOLLS[1,a], JEONG EUN LEE[2,b], NICHOLAS KARN[3,c], DAVID JOHNSON[4,d], RUKUANG HUANG[5,e] AND ALEXIS MUIR-WATT[6,f]

[1]*Department of Statistics, University of Oxford,* [a]*nicholls@stats.ox.ac.uk*
[2]*Department of Statistics, University of Auckland,* [b]*kate.lee@auckland.ac.nz*
[3]*Faculty of Arts and Humanities, University of Southampton,* [c]*N.E.Karn@soton.ac.uk*
[4]*St Peter's college, University of Oxford,* [d]*david.johnson@spc.ox.ac.uk*
[5]*Department of Psychiatry, University of Oxford,* [e]*rukuang.huang@jesus.ox.ac.uk*
[6]*Department of Statistics, University of Oxford,* [f]*alexis.muirwatt@gmail.com*

In the eleventh and twelfth centuries in England, Wales and Normandy, royal acta were legal documents in which witnesses were listed in order of social status. Any bishops present were listed as a group. For our purposes each witness-list is an ordered permutation of bishop names with a known date or date-range. Changes over time in the order bishops are listed may reflect changes in their authority. Historians would like to detect and quantify these changes. There is no reason to assume that the underlying social order, which constrains bishop-order, within lists is a complete order. We therefore model the evolving social order as an evolving partial ordered set or *poset*.

We construct a hidden Markov model for these data. The hidden state is an evolving poset (the evolving social hierarchy) and the emitted data are random total orders (dated lists) respecting the poset present at the time the order was observed. This generalises existing models for rank-order data such as Mallows and Plackett–Luce. We account for noise via a random "queue-jumping" process. Our latent-variable prior for the random process of posets is marginally consistent. A parameter controls poset depth, and actor-covariates inform the position of actors in the hierarchy. We fit the model, estimate posets and find evidence for changes in status over time. We interpret our results in terms of court politics. Simpler models, based on bucket orders and vertex-series-parallel orders, are rejected. We compare our results with a time-series extension of the Plackett–Luce model. Our software is publicly available.

REFERENCES

ARCAGNI, A., AVELLONE, A. and FATTORE, M. (2022). Complexity reduction and approximation of multidomain systems of partially ordered data. *Comput. Statist. Data Anal.* **173** Paper No. 107520, 23. MR4418926 https://doi.org/10.1016/j.csda.2022.107520

ASFAW, D., VITELLI, V., SØRENSEN, Ø., ARJAS, E. and FIRGESSI, A. (2017). Time-varying rankings with the Bayesian Mallows model. *Stat* **6** 14–30. MR3599889 https://doi.org/10.1002/sta4.132

BATES, D. (2013). *The Normans and Empire*. *Ford Lectures*. OUP, Oxford.

BEERENWINKEL, N., ERIKSSON, N. and STURMFELS, B. (2007). Conjunctive Bayesian networks. *Bernoulli* **13** 893–909. MR2364218 https://doi.org/10.3150/07-BEJ6133

BOGART, K. P. (1973a). Maximal dimensional partially ordered sets. I. Hiraguchi's theorem. *Discrete Math.* **5** 21–31. MR0318013 https://doi.org/10.1016/0012-365X(73)90024-1

BOGART, K. P. (1973b). Preference structures. I. Distances between transitive preference relations. *J. Math. Sociol.* **3** 49–67. MR0395936 https://doi.org/10.1080/0022250x.1973.9989823

BRIGHTWELL, G. (1993). Models of random partial orders. In *Surveys in Combinatorics*. *London Mathematical Society Lecture Note Series* **187** 53–83. Cambridge Univ. Press, Cambridge. MR1239232 https://doi.org/10.1017/CBO9780511662089.004

Key words and phrases. Partial order, Bayesian analysis, hidden Markov model, royal acta, social hierarchy.

BRIGHTWELL, G. and WINKLER, P. (1991). Counting linear extensions. *Order* **8** 225–242. MR1154926 https://doi.org/10.1007/BF00383444

CARON, F. and DOUCET, A. (2012). Efficient Bayesian inference for generalized Bradley–Terry models. *J. Comput. Graph. Statist.* **21** 174–196. MR2913362 https://doi.org/10.1080/10618600.2012.638220

CARON, F. and TEH, Y. (2012). Bayesian nonparametric models for ranked data. In *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou and K. Weinberger, eds.) **25**. Curran Associates, Red Hook.

CARON, F., TEH, Y. W. and MURPHY, T. B. (2014). Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. *Ann. Appl. Stat.* **8** 1145–1181. MR3262549 https://doi.org/10.1214/14-AOAS717

CLOVER, H. and GIBSON, M. (1979). *The Letters of Lanfranc, Archbishop of Canterbury*. Oxford Univ. Press, London.

DENG, K., HAN, S., LI, K. J. and LIU, J. S. (2014). Bayesian aggregation of order-based rank data. *J. Amer. Statist. Assoc.* **109** 1023–1039. MR3265678 https://doi.org/10.1080/01621459.2013.878660

DIACONIS, P. (1988). *Group Representations in Probability and Statistics. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **11**. IMS, Hayward, CA. MR0964069

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

DURANTE, D., DUNSON, D. B. and VOGELSTEIN, J. T. (2017b). Rejoinder: Nonparametric Bayes modeling of populations of networks. *J. Amer. Statist. Assoc.* **112** 1547–1552. MR3750879 https://doi.org/10.1080/01621459.2017.1395643

FEA (2022). *Fasti Ecclesiae Anglicanae* 1066–1300. London, Institute of Historical Research. British History Online. Last accessed 22 October 2022. Available at https://www.british-history.ac.uk/series/fasti-ecclesiae-anglicanae.

FISHBURN, P. C. and GEHRLEIN, W. V. (1975). A comparative analysis of methods for constructing weak orders from partial orders. *J. Math. Sociol.* **4** 93–102. MR0406580 https://doi.org/10.1080/0022250x.1975.9989846

FLIGNER, M. A. and VERDUCCI, J. S. (1986). Distance based ranking models. *J. Roy. Statist. Soc. Ser. B* **48** 359–369. MR0876847

FOERSTER, S., FRANZ, M., MURRAY, C. M., GILBY, I. C., FELDBLUM, J. T., WALKER, K. K. and PUSEY, A. E. (2016). Chimpanzee females queue but males compete for social status. *Sci. Rep.* **6** 35404–35415.

FRIEDELL, M. F. (1967). Organizations as semilattices. *Amer. Sociol. Rev.* **32** 46–54.

FROEHLICH, H., FELLMANN, M., SUELTMANN, H., POUSTKA, A. and BEISSBARTH, T. (2007). Large scale statistical inference of signaling pathways from RNAi and microarray data. *BMC Bioinform.* **8** 386–390.

GAZES, R. P., HAMPTON, R. R. and LOURENCO, S. F. (2017). Transitive inference of social dominance by human infants. *Dev. Sci.* **20** e12367.

GIONIS, A., MANNILA, H., PUOLAMÄKI, K. and UKKONEN, A. (2006). Algorithms for discovering bucket orders from data. In *Proceedings of the* 12*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 561–566.

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

GUIVER, J. and SNELSON, E. (2009). Bayesian inference for Plackett–Luce ranking models. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 377–384.

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.

HASKINS, G. L. (1938). Charter witness lists in the reign of King John. *Speculum* **13** 319–325.

HIRAGUCHI, T. (1951). On the dimension of partially ordered sets. *Sci. Rep. Kanazawa Univ.* **1** 77–94. MR0070681

HUNTER, D. R. (2004). MM algorithms for generalized Bradley–Terry models. *Ann. Statist.* **32** 384–406. MR2051012 https://doi.org/10.1214/aos/1079120141

IRUROZKI, E., CALVO, B. and LOZANO, J. A. (2016). PerMallows: An R package for Mallows and generalized Mallows models. *J. Stat. Softw.* **71** 1–30.

IRUROZKI, E., CALVO, B. and LOZANO, J. A. (2019). Mallows and generalized Mallows model for matchings. *Bernoulli* **25** 1160–1188. MR3920369 https://doi.org/10.3150/17-bej1017

JIANG, C. and NICHOLLS, G. K. (2024). Non-parametric Bayesian inference for partial orders with ties from rank data observed with Mallows noise. Available at arXiv:2408.14661.

JIANG, C., NICHOLLS, G. K. and LEE, J. E. (2023). Bayesian inference for vertex-series-parallel partial orders. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence* (R. J. Evans and I. Shpitser, eds.). *Proceedings of Machine Learning Research* **216** 995–1004. PMLR, Cambridge MA.

JOHNSON, D. (2013). Bishops and deans: London and the province of Canterbury in the twelfth century. *Hist*. *Res*. **86** 551–578.

KANGAS, K., HANKALA, T., NIINIMÄKI, T. and KOIVISTO, M. (2016). Counting linear extensions of sparse posets. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, *IJCAI*'16 603–609. AAAI Press, Menlo Park.

KANGAS, K., KOIVISTO, M. and SALONEN, S. (2020). A faster tree-decomposition based algorithm for counting linear extensions. *Algorithmica* **82** 2156–2173. MR4132887 https://doi.org/10.1007/s00453-019-00633-1

KARZANOV, A. and KHACHIYAN, L. (1991). On the conductance of order Markov chains. *Order* **8** 7–15. MR1129609 https://doi.org/10.1007/BF00385809

KLEITMAN, D. J. and ROTHSCHILD, B. L. (1975). Asymptotic enumeration of partial orders on a finite set. *Trans*. *Amer*. *Math*. *Soc*. **205** 205–220. MR0369090 https://doi.org/10.2307/1997200

KNUTH, D. E. and SZWARCFITER, J. L. (1974). A structured program to generate all topological sorting arrangements. *Inform*. *Process*. *Lett*. **2** 153–157.

LU, T. and BOUTILIER, C. (2014). Effective sampling and learning for Mallows models with pairwise-preference data. *J*. *Mach*. *Learn*. *Res*. **15** 3783–3829. MR3317212

LUCE, R. D. (1959). On the possible psychophysical laws. *Psychol*. *Rev*. **66** 81.

MALLOWS, C. L. (1957). Non-null ranking models. I. *Biometrika* **44** 114–130. MR0087267 https://doi.org/10.1093/biomet/44.1-2.114

MANNILA, H. (2008). Finding total and partial orders from data for seriation. In *Discovery Science* (J.-F. Boucault, ed.). *LNAI* **5255** 16–25. Springer, Berlin Heidelberg.

MANNILA, H. and MEEK, C. (2000). Global partial orders from sequential data. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *KDD*'00, *New York*, *NY*, *USA* 161–168. Assoc. Comput. Mach., New York.

MARTIN, J. L. (2002). Some algebraic structures for diffusion in social networks. *J*. *Math*. *Sociol*. **26** 123–146.

MCKEOUGH, K. and GLICKMAN, M. (2024). Plackett–Luce modeling with trajectory models for measuring athlete strength. *J*. *Quant*. *Anal*. *Sports* **20** 21–35.

MEILĂ, M. and CHEN, H. (2010). Dirichlet process mixtures of generalized Mallows models. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, *UAI*'10, *Arlington*, *Virginia*, *USA* 358–367. AUAI Press, Arlington, VA.

MEILĂ, M. and CHEN, H. (2016). Bayesian non-parametric clustering of ranking data. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **38** 2156–2169.

MOGAPI, O. (2009). A Latent Partial Order Model for Social Networks. Ph. D. thesis, Univ. Oxford.

MOLLICA, C. and TARDELLA, L. (2017). Bayesian Plackett–Luce mixture models for partially ranked data. *Psychometrika* **82** 442–458. MR3665302 https://doi.org/10.1007/s11336-016-9530-0

MOLLICA, C. and TARDELLA, L. (2020). PLMIX: An R package for modelling and clustering partially ranked data. *J*. *Stat*. *Comput*. *Simul*. **90** 925–959. MR4067354 https://doi.org/10.1080/00949655.2020.1711909

MUIR WATT, A. (2015). Inference for partial orders from random linear extensions. Ph. D. thesis, Univ. of Oxford.

NICHOLLS, G. K., LEE, J. E., KARN, N., JOHNSON, D., HUANG, R. and MUIR-WATT, A. (2025). Supplement to "Bayesian inference for partial orders from random linear extensions: power relations from 12th Century Royal Acta." https://doi.org/10.1214/24-AOAS2002SUPPA, https://doi.org/10.1214/24-AOAS2002SUPPB

NICHOLLS, G. K. and MUIR WATT, A. (2011). Partial order models for episcopal social status in 12th century England. In *IWSM* **2011** 437.

NIINIMÄKI, T., PARVIAINEN, P. and KOIVISTO, M. (2016). Structure discovery in Bayesian networks by sampling partial orders. *J*. *Mach*. *Learn*. *Res*. **17** Paper No. 57, 47. MR3504617

OEIS FOUNDATION INC. (2022). The on-line encyclopedia of integer sequences. Available at https://oeis.org.

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

RISING, J. (2021). Uncertainty in ranking. Available at arXiv:2107.03459.

ROBERTS, J. M. (1990). Modeling hierarchy: Transitivity and the linear ordering problem. *J*. *Math*. *Sociol*. **16** 77–87.

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

RUSSELL, J. C. (1937). Social status at the court of King John. *Speculum* **12** 319–329.

SAKOPARNIG, T. and BEERENWINKEL, N. (2012). Efficient sampling for Bayesian inference of conjunctive Bayesian networks. *Bioinformatics* **28** 2318–2324.

SESHADRI, A., RAGAIN, S. and UGANDER, J. (2020). Learning rich rankings. In *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin, eds.) **33** 9435–9446. Curran Associates, Red Hook.

SHARPE, R., CARPENTER, D., DOHERTY, H., HAGGER, M. and KARN, N. (2014). The charters of William II and Henry I. Accessed October 2022. Available at actswilliam2henry1.wordpress.com/the-charters/.

SHIZUKA, D. and MCDONALD, D. B. (2012). A social network perspective on measurements of dominance hierarchies. *Anim. Behav.* **83** 925–934.

SØRENSEN, Ø., CRISPINO, M., LIU, Q. and VITELLI, V. (2020). BayesMallows: An R package for the Bayesian Mallows model. *R J.* **12** 324–342.

TKACHENKO, M. and LAUW, H. W. (2016). Plackett–Luce regression mixture model for heterogeneous rankings. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* 237–246.

VALDES, J. (1978). Parsing Flowcharts and Series-Parallel Graphs. Ph. D. thesis, Stanford, CA, USA. AAI7905944.

VALDES, J., TARJAN, R. E. and LAWLER, E. L. (1982). The recognition of series parallel digraphs. *SIAM J. Comput.* **11** 298–313. MR0652904 https://doi.org/10.1137/0211023

VAN WIETMARSCHEN, H. (2022). What is social hierarchy? *Noûs* **56** 920–939. (page 926, “In sum, . . . ”).

VASCONCELOS, M. (2008). Transitive inference in non-human animals: An empirical and theoretical analysis. *Behav. Process.* **78** 313–334.

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

VITELLI, V., SØRENSEN, Ø., CRISPINO, M., FRIGESSI, A. and ARJAS, E. (2018). Probabilistic preference learning with the Mallows rank model. *J. Mach. Learn. Res.* **18** Paper No. 158, 49. MR3813807

WELLS, M. B. (1971). *Elements of Combinatorial Computing*. Pergamon, Oxford. MR0277082

WINKLER, P. (1985). Random orders. *Order* **1** 317–331. MR0787544 https://doi.org/10.1007/BF00582738

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1691–1713
https://doi.org/10.1214/24-AOAS2003

ESTIMATING REPORTING BIAS IN 311 COMPLAINT DATA

BY KATE S. BOXER[1,a], BOYEONG HONG[2,c], CONSTANTINE E. KONTOKOSTA[2,d] AND DANIEL B. NEILL[1,b]

[1]*Machine Learning for Good Laboratory, Center for Urban Science and Progress, New York University,* [a]*kb145@nyu.edu,* [b]*daniel.neill@nyu.edu*

[2]*Marron Institute of Urban Management, New York University,* [c]*bh1555@nyu.edu,* [d]*ckontokosta@nyu.edu*

Systems such as "311" enable residents of a community to report on their environments and to request nonemergency municipal services. While such systems provide an important link between community and government, resident-generated data suffer from reporting bias, with some subpopulations reporting at lower rates than others. Our research focuses on defining the underreporting of heating and hot water problems to New York City's 311 system and developing methods to estimate under-reporting. First, we estimate nonreporting by fitting a latent variable model, which estimates both the probability of an underlying heating problem conditional on building characteristics, and the probability of reporting a problem conditional on population characteristics. Second, we analyze "less-than-expected" reporting: buildings with fewer 311 calls than expected, as compared to similarly-sized buildings with similar estimated problem durations. Together, these analyses determine neighborhoods and neighborhood-level socioeconomic characteristics that are predictive of underreporting of heating and hot water problems. Our approaches can aid government agencies wishing to use resident-generated data to assist in constructing fair public policies.

REFERENCES

BEKKER, J. and DAVIS, J. (2020). Learning from positive and unlabeled data: A survey. *Mach. Learn.* **109** 719–760. MR4094532 https://doi.org/10.1007/s10994-020-05877-5

BEKKER, J., ROBBERECHTS, P. and DAVIS, J. (2019). Beyond the selected completely at random assumption for learning from positive and unlabeled data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 71–85. Springer, Berlin.

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

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

BOXER, K. S., HONG, B., KONTOKOSTA, C. E. and NEILL, D. B. (2025). Supplement to "Estimating reporting bias in 311 complaint data." https://doi.org/10.1214/24-AOAS2003SUPP

CAUCHY, A. et al. (1847). Méthode générale pour la résolution des systemes d'équations simultanées. *C. R. Sci. Paris* **25** 536–538.

CITY OF NEW ORLEANS (2019). Catch basins. Available at https://nola.gov/dpw/catch-basins/. Accessed 2021-10-05.

CITY OF NEW YORK (2021). About NYC311. Available at https://portal.311.nyc.gov/about-nyc-311/.

DARITY JR., W., HAMILTON, D., PAUL, M., AJA, A., PRICE, A., MOORE, A. and CHIOPRIS, C. (2018). What we get wrong about closing the racial wealth gap. *Samuel DuBois Cook Center on Social Equity and Insight Center for Community Economic Development* **1** 1–67.

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

ELKAN, C. and NOTO, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 213–220.

KATO, M., TESHIMA, T. and HONDA, J. (2019). Learning from positive and unlabeled data with a selection bias. In *International Conference on Learning Representations*.

Key words and phrases. Resident-generated data, citizen-generated data, 311 data, city analytics, latent variable models, reporting bias, positive and unlabeled learning.

KELLY, J. M. and SWINDELL, D. (2002). Service quality variation across urban space: First steps toward a model of citizen satisfaction. *J. Urban Aff.* **24** 271–288.

KONTOKOSTA, C. E. (2016). The quantified community and neighborhood labs: A framework for computational urban science and civic technology innovation. *J. Urban Technol.* **23** 67–84.

KONTOKOSTA, C. E. and HONG, B. (2021). Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions. *Sustain. Cities Soc.* **64** 102503.

KONTOKOSTA, C. E., WEISS, M., SNIVELY, C. and GULICK, S. (2017). NYC311. Harvard Business School Case 818-056.

MASSEY JR, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *J. Amer. Statist. Assoc.* **46** 68–78.

MCLAFFERTY, S., SCHNEIDER, D. and ABELT, K. (2020). Placing volunteered geographic health information: Socio-spatial bias in 311 bed bug report data for New York City. *Health Place* **62** 102282. https://doi.org/10.1016/j.healthplace.2019.102282

MEIJER, A. and POTJER, S. (2018). Citizen-generated open data: An explorative analysis of 25 cases. *Gov. Inf. Q.* **35** 613–621.

MINKOFF, S. L. (2016). NYC 311: A tract-level analysis of citizen–government contacting in New York City. *Urban Aff. Rev.* **52** 211–246.

NADEAU, L. K. (2011). New York City unveils real-time 311 request map. Available at http://www.govtech.com/e-government/New-York-City-Real-Time-311-Map-021711.html. Accessed 2017-07-08.

RAMASWAMY, H., SCOTT, C. and TEWARI, A. (2016). Mixture proportion estimation via kernel embeddings of distributions. In *International Conference on Machine Learning* 2052–2060. PMLR.

WARD, G., HASTIE, T., BARRY, S., ELITH, J. and LEATHWICK, J. R. (2009). Presence-only data and the EM algorithm. *Biometrics* **65** 554–563. MR2751480 https://doi.org/10.1111/j.1541-0420.2008.01116.x

WHITE, A. and TRUMP, K.-S. (2018). The promises and pitfalls of 311 data. *Urban Aff. Rev.* **54** 794–823.

ZHA, Y. and VELOSO, M. (2014). Profiling and prediction of non-emergency calls in New York City. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

ZHENG, Y., LIU, T., WANG, Y., ZHU, Y., LIU, Y. and CHANG, E. (2014). Diagnosing New York city's noises with ubiquitous data. In *Proceedings of the* 2014 *ACM International Joint Conference on Pervasive and Ubiquitous Computing* 715–725. ACM, New York.

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1714–1733
https://doi.org/10.1214/24-AOAS2004

FLEXIBLE BAYESIAN SPATIAL MODELING FOR UNKNOWN MISSING DATA MECHANISM IN SURVEY ANALYSIS: AN APPLICATION TO THE CHINESE GENERAL SOCIETY SURVEY

BY GUANYU HU[1,a], MING-HUI CHEN[2,b] AND ZHIHUA MA[3,c]

[1]*Center for Spatial Temporal Modeling for Applications in Population Sciences, Department of Biostatistics and Data Science, The University of Texas Health Science Center at Houston,* [a]*guanyu.hu@uth.tmc.com*

[2]*Department of Statistics, University of Connecticut,* [b]*ming-hui.chen@uconn.edu*

[3]*College of Economics, Shenzhen University,* [c]*mazh1993@outlook.com*

Social survey data are often collected from different survey centers in different regions. In some circumstances the response variables are completely observed while the covariates have missing values. In addition, the missing data patterns will vary in different geographical locations. In this article an ordinal response regression model with a log linear model for the cutpointsis extended to fit the spatial ordinal response data with missing covariates within a Bayesian framework. A signed-spike-and-slab prior is developed to learn the heterogeneity of the missing data mechanisms among different spatial locations. The properties of the proposed models are examined, and a Markov chain Monte Carlo sampling algorithm is used to sample from the posterior distribution. Extensive simulation studies are carried out to examine the empirical performance of the proposed methods. We further apply the proposed methodology to analyze a real dataset from a Chinese General Society Survey.

REFERENCES

ARNONE, E., SANGALLI, L. M. and VICINI, A. (2023). Smoothing spatio-temporal data with complex missing data patterns. *Stat. Model.* **23** 327–356. MR4624327 https://doi.org/10.1177/1471082X211057959

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

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

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

BRICK, J. M. and KALTON, G. (1996). Handling missing data in survey research. *Stat. Methods Med. Res.* **5** 215–238. https://doi.org/10.1177/096228029600500302

BROOK, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika* **51** 481–483. MR0205315 https://doi.org/10.1093/biomet/51.3-4.481

BURT, R. S. (1987). A note on missing network data in the general social survey. *Soc. Netw.* **9** 63–73.

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

FERRER-I CARBONELL, A. and RAMOS, X. (2014). Inequality and happiness. *J. Econ. Surv.* **28** 1016–1027.

FOLLMANN, D. and WU, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51** 151–168. MR1341233 https://doi.org/10.2307/2533322

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

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

Key words and phrases. Conditional autoregression, nonignorable missingness, personal income, spatial heterogeneity.

GWON, Y., MO, M., CHEN, M.-H., CHI, Z., LI, J., XIA, A. H. and IBRAHIM, J. G. (2020). Network meta-regression for ordinal outcomes: Applications in comparing Crohn's disease treatments. *Stat*. *Med*. **39** 1846–1870. MR4098525 https://doi.org/10.1002/sim.8518

HORDIJK, L. (1974). Spatial correlation in the disturbances of a linear interregional model. *Reg*. *Urban Econ*. **4** 117–140.

HU, G., CHEN, M.-H. and MA, Z. (2025). Supplement to "Flexible Bayesian Spatial Modeling for Unknown Missing Data Mechanism in Survey Analysis: an Application to the Chinese General Society Survey." https://doi.org/10.1214/24-AOAS2004SUPPA, https://doi.org/10.1214/24-AOAS2004SUPPB

HU, G., XUE, Y. and MA, Z. (2023). Bayesian clustered coefficients regression with auxiliary covariates assistant random effects. *Stat*. *Model*. **23** 273–293. MR4594787 https://doi.org/10.1177/1471082X211049278

HUANG, L., CHEN, M.-H. and IBRAHIM, J. G. (2005). Bayesian analysis for generalized linear models with nonignorably missing covariates. *Biometrics* **61** 767–780. MR2196165 https://doi.org/10.1111/j.1541-0420.2005.00338.x

IBRAHIM, J. G., LIPSITZ, S. R. and CHEN, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **61** 173–190. MR1664045 https://doi.org/10.1111/1467-9868.00170

JEBB, A. T., TAY, L., DIENER, E. and OISHI, S. (2018). Happiness, income satiation and turning points around the world. *Nat*. *Hum*. *Behav*. **2** 33–38. https://doi.org/10.1038/s41562-017-0277-0

KNIGHT, J., LINA, S. and GUNATILAKA, R. (2009). Subjective well-being and its determinants in rural China. *China Econ*. *Rev*. **20** 635–649.

LEÓN-NOVELO, L. G. and SAVITSKY, T. D. (2019). Fully Bayesian estimation under informative sampling. *Electron*. *J*. *Stat*. **13** 1608–1645. MR3939589 https://doi.org/10.1214/19-ejs1538

LIM, D. K., RASHID, N. U., OLIVA, J. B. and IBRAHIM, J. G. (2021). Handling non-ignorably missing features in electronic health records data using importance-weighted autoencoders. arXiv preprint. Available at arXiv:2101.07357.

LITTLE, R. and RUBIN, D. (2019). *Statistical Analysis with Missing Data*, *Third Edition*. Wiley, New York.

MA, Z., HU, G. and CHEN, M.-H. (2021). Bayesian hierarchical spatial regression models for spatial data in the presence of missing covariates with applications. *Appl*. *Stoch*. *Models Bus*. *Ind*. **37** 342–359. MR4247420 https://doi.org/10.1002/asmb.2568

MASON, A., RICHARDSON, S. and BEST, N. (2012). Two-pronged strategy for using DIC to compare selection models with non-ignorable missing responses. *Bayesian Anal*. **7** 109–146. MR2896714 https://doi.org/10.1214/12-BA704

MENA, R. H. and WALKER, S. G. (2015). On the Bayesian mixture model and identifiability. *J*. *Comput*. *Graph*. *Statist*. **24** 1155–1169. MR3432934 https://doi.org/10.1080/10618600.2014.950376

MIAO, W., DING, P. and GENG, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *J*. *Amer*. *Statist*. *Assoc*. **111** 1673–1683. MR3601726 https://doi.org/10.1080/01621459.2015.1105808

MILITINO, A. F., UGARTE, M. and MONTESINO, M. (2019). Filling missing data and smoothing altered data in satellite imagery with a spatial functional procedure. *Stoch*. *Environ*. *Res*. *Risk Assess*. **33** 1737–1750.

MORIKAWA, K. and KIM, J. K. (2021). Semiparametric optimal estimation with nonignorable nonresponse data. *Ann*. *Statist*. **49** 2991–3014. MR4338901 https://doi.org/10.1214/21-aos2070

NSRC (2018). China general social survey. http://cgss.ruc.edu.cn/.

PADILLA, L., LAGOS-ÁLVAREZ, B., MATEU, J. and PORCU, E. (2020). Space-time autoregressive estimation and prediction with missing data based on Kalman filtering. *Environmetrics* **31** e2627. MR4166853 https://doi.org/10.1002/env.2627

QIN, K., LUO, P., LU, B. and LIN, Z. (2020). Analyzing spatial patterns of intergenerational education mobility in China. In *Spatial Synthesis* 285–301. Springer, Berlin.

ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J*. *Amer*. *Statist*. *Assoc*. **90** 122–129. MR1325119

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Classics Library*. Wiley-Interscience, Hoboken, NJ. MR2117498

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

SMITH, T. W. (1991). *An Analysis of Missing Income Information on the General Social Surveys*. NORC, Univ. Chicago, Chicago.

SU, Y.-S., LIEN, D. and YAO, Y. (2022). Economic growth and happiness in China: A Bayesian multilevel age-period-cohort analysis based on the CGSS data 2005–2015. *Int*. *Rev*. *Econ*. *Finance* **77** 191–205.

SUGASAWA, S., MORIKAWA, K. and TAKAHATA, K. (2022). Bayesian semiparametric modeling of response mechanism for nonignorable missing data. *TEST* **31** 101–117. MR4386428 https://doi.org/10.1007/s11749-021-00774-y

YAO, H., KIM, S., CHEN, M.-H., IBRAHIM, J. G., SHAH, A. K. and LIN, J. (2015). Bayesian inference for multivariate meta-regression with a partially observed within-study sample covariance matrix. *J. Amer. Statist. Assoc*. **110** 528–544. MR3367245 https://doi.org/10.1080/01621459.2015.1006065

ZHANG, D., CHEN, M.-H., IBRAHIM, J. G., BOYE, M. E., WANG, P. and SHEN, W. (2014). Assessing model fit in joint models of longitudinal and survival data with applications to cancer clinical trials. *Stat. Med*. **33** 4715–4733. MR3274507 https://doi.org/10.1002/sim.6269

ZHOU, J., LIU, J., WANG, F. and WANG, H. (2022). Autoregressive model with spatial dependence and missing data. *J. Bus. Econom. Statist*. **40** 28–34. MR4356555 https://doi.org/10.1080/07350015.2020.1766471

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1734–1752
https://doi.org/10.1214/24-AOAS2008

SUPERVISED CENTRALITY VIA SPARSE NETWORK INFLUENCE REGRESSION: AN APPLICATION TO THE 2021 HENAN FLOODS' SOCIAL NETWORK

BY YINGYING MA[1,a], WEI LAN[2,b], CHENLEI LENG[3,c], TING LI[4,d] AND HANSHENG WANG[5,e]

[1]*School of Economics and Management, Beihang University,* [a]*mayingying@buaa.edu.cn*
[2]*Center of Statistical Research, Southwestern University of Finance and Economics,* [b]*lanwei@swufe.edu.cn*
[3]*Department of Statistics, University of Warwick,* [c]*C.Leng@warwick.ac.uk*
[4]*Department of Applied Mathematics, Hong Kong Polytechnic University,* [d]*tingeric.li@polyu.edu.hk*
[5]*Guanghua School of Management, Peking University,* [e]*hansheng@pku.edu.cn*

The social characteristics of players in a social network are closely associated with their network positions and relational importance. Identifying those influential players in a network is of great importance, as it helps to understand how ties are formed, how information is propagated, and, in turn, can guide the dissemination of new information. Motivated by a Sina Weibo social network analysis of the 2021 Henan Floods, where response variables for each Sina Weibo user are available, we propose a new notion of supervised centrality that emphasizes the task-specific nature of a player's centrality. To estimate the supervised centrality and identify important players, we develop a novel sparse network influence regression by introducing individual heterogeneity for each user. To overcome the computational difficulties in fitting the model for large social networks, we further develop a forward-addition algorithm and show that it can consistently identify a superset of the influential Sina Weibo users. We apply our method to analyze three responses in the Henan Floods data: the number of comments, reposts, and likes, and obtain meaningful results. A further simulation study corroborates the developed method.

REFERENCES

ADAMIC, L. A. and HUBERMAN, B. A. (2000). Power-law distribution of the world wide web. *Science* **287** 2115–2115.

CENTOLA, D. (2010). The spread of behavior in an online social network experiment. *Science* **329** 1194–1197.

CERCEL, D.-C. and TRAUSAN-MATU, S. (2014). Opinion propagation in online social networks: A survey. In *Proceedings of the* 4*th International Conference on Web Intelligence*, *Mining and Semantics* (*WIMS*14) 1–10.

CHA, M., HADDADI, H., BENEVENUTO, F. and GUMMADI, K. (2010). Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the International AAAI Conference on Web and Social Media* **4** 10–17.

CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189 https://doi.org/10.1093/biomet/asn034

CLIFF, A. D. and ORD, J. K. (1975). Model building and the analysis of spatial pattern in human geography. *J. Roy. Statist. Soc. Ser. B* **37** 297–348. MR0413415

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

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

FREEMAN, L. C. (1978). Centrality in social networks conceptual clarification. *Soc. Netw.* **1** 215–239.

FREEMAN, L. C. et al. (2002). Centrality in social networks: Conceptual clarification. In *Social Network*: *Critical Concepts in Sociology* **1** 238–263. Londres, Routledge.

Key words and phrases. Screening, social networks, spatial autoregression, supervised centrality.

GOLBECK, J. and HENDLER, J. (2006). Inferring binary trust relationships in web-based social networks. *ACM Trans. Internet Technol.* **6** 497–529.

GRUHL, D., GUHA, R., LIBEN-NOWELL, D. and TOMKINS, A. (2004). Information diffusion through blogspace. In *Proceedings of the* 13*th International Conference on World Wide Web* 491–501.

GUO, S., WANG, Y. and YAO, Q. (2016). High-dimensional and banded vector autoregressions. *Biometrika* **103** 889–903. MR3620446 https://doi.org/10.1093/biomet/asw046

HANSEN, D. L., SHNEIDERMAN, B., SMITH, M. A. and HIMELBOIM, I. (2020). Chapter 3—social network analysis: Measuring, mapping, and modeling collections of connections. In *Analyzing Social Media Networks with NodeXL* (*Second Edition*), 2nd ed. 31–51. Morgan Kaufmann, San Mateo.

JIN, B., ZOU, M., WEI, Z. and GUO, W. (2024). How to find opinion leader on the online social network? *J. Netw. Comput. Appl.* **00** 1–15.

KEMPE, D., KLEINBERG, J. and TARDOS, É. (2005). Influential nodes in a diffusion model for social networks. In *Automata, Languages and Programming. Lecture Notes in Computer Science* **3580** 1127–1138. Springer, Berlin. MR2184706 https://doi.org/10.1007/11523468_91

LEENDERS, R. T. A. J. (2002). Modeling social influence through network autocorrelation: Constructing the weight matrix. *Soc. Netw.* **24** 21–47.

LEWBEL, A., QU, X. and TANG, X. (2022). Social networks with unobserved links. *J. Polit. Econ.*

LIN, C. T. (1974). Structural controllability. *IEEE Trans. Automat. Control* **AC-19** 201–208. MR0452870 https://doi.org/10.1109/tac.1974.1100557

LIN, X. (2010). Identifying peer effects in student academic achievement by spatial autoregressive models with group unobservables. *J. Labor Econ.* **28** 825–860.

LIU, X. (2013). Estimation of local-aggregate network model with sampled networks. *Econom. Lett.* **118** 243–246. MR3001465 https://doi.org/10.1016/j.econlet.2012.10.037

LIU, X. and LEE, L. (2010). GMM estimation of social interaction models with centrality. *J. Econometrics* **159** 99–115. MR2720846 https://doi.org/10.1016/j.jeconom.2010.04.009

LIU, X., PATACCHINI, E. and ZENOU, Y. (2014). Endogenous peer effects: Local aggregate or local average? *J. Econ. Behav. Organ.* **103** 39–59.

MA, Y., GUO, S. and WANG, H. (2023). Sparse spatio-temporal autoregressions by profiling and bagging. *J. Econometrics* **232** 132–147. MR4513104 https://doi.org/10.1016/j.jeconom.2020.10.010

MA, Y., LAN, W., LENG, C., LI, T. and WANG, H. (2025). Supplement to "Supervised Centrality via Sparse Network Influence Regression: an Application to the 2021 Henan Floods' Social Network." https://doi.org/10.1214/24-AOAS2008SUPPA, https://doi.org/10.1214/24-AOAS2008SUPPB

MARCHIORI, M. and LATORA, V. (2000). Harmony in the small-world. *Phys. A, Stat. Mech. Appl.* **285** 539–546.

MOTTER, A. E. and LAI, Y. C. (2002). Cascade-based attacks on complex networks. *Phys. Rev. E* **66** 65102–65102.

NEWMAN, M. E. J. (2010). *Networks: An Introduction*. Oxford Univ. Press, Oxford. MR2676073 https://doi.org/10.1093/acprof:oso/9780199206650.001.0001

NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96** 1077–1087. MR1947255 https://doi.org/10.1198/016214501753208735

ORD, K. (1975). Estimation methods for models of spatial interaction. *J. Amer. Statist. Assoc.* **70** 120–126. MR0413393

QU, X., LEE, L. and YANG, C. (2021). Estimation of a SAR model with endogenous spatial weights constructed by bilateral variables. *J. Econometrics* **221** 180–197. MR4208971 https://doi.org/10.1016/j.jeconom.2020.05.011

SIERSDORFER, S., CHELARU, S., PEDRO, J. S., ALTINGOVDE, I. S. and NEJDL, W. (2014). Analyzing and mining comments and comment ratings on the social web. *ACM Trans. Web* **8** 17.

SONG, X., CHI, Y., HINO, K. and TSENG, B. (2007). Identifying opinion leaders in the blogosphere. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* 971–974.

SUN, J. and TANG, J. (2011). A survey of models and algorithms for social influence analysis. In *Social Network Data Analytics* 177–214. Springer, New York. MR3014049 https://doi.org/10.1007/978-1-4419-8462-3_7

VICARIO, M. D., QUATTROCIOCCHI, W., SCALA, A. and ZOLLO, F. (2019). Polarization and fake news: Early warning of potential misinformation targets. *ACM Trans. Web* **13** 1–22.

WANG, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.* **104** 1512–1524. MR2750576 https://doi.org/10.1198/jasa.2008.tm08516

WASSERMAN, S. and FAUST, K. (1994). *Social Network Analysis: Methods and Applications*.

WEI, Z., PAGANI, A., FU, G., GUYMER, I., CHEN, W., MCCANN, J. and GUO, W. (2020). Optimal sampling of water distribution network dynamics using graph Fourier transform. *IEEE Trans. Netw. Sci. Eng.* **7** 1570–1582. MR4150024 https://doi.org/10.1109/tnse.2019.2941834

YANG, L., QIAO, Y., LIU, Z., MA, J. and LI, X. (2018). Identifying opinion leader nodes in online social networks with a new closeness evaluation algorithm. *Soft Comput.* **22** 453–464.

ZHAO, Y., WU, Y.-J., LEVINA, E. and ZHU, J. (2017). Link prediction for partially observed networks. *J. Comput. Graph. Statist.* **26** 725–733. MR3698680 https://doi.org/10.1080/10618600.2017.1286243

ZHOU, J., TU, Y., CHEN, Y. and WANG, H. (2017). Estimating spatial autocorrelation with sampled network data. *J. Bus. Econom. Statist.* **35** 130–138. MR3591541 https://doi.org/10.1080/07350015.2015.1061437

ZHU, X., CHANG, X., LI, R. and WANG, H. (2019). Portal nodes screening for large scale social networks. *J. Econometrics* **209** 145–157. MR3944747 https://doi.org/10.1016/j.jeconom.2018.12.021

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
2025, Vol. 19, No. 2, 1753–1777
https://doi.org/10.1214/25-AOAS2020

PREDICTING GENDER EMPLOYMENT DISCREPANCIES: A MULTIVARIATE FAY-HERRIOT MODEL FOR TRANSFORMED PROPORTIONS

BY ESTEBAN CABELLO[1,a], DOMINGO MORALES[1,b] AND AGUSTÍN PÉREZ[2,c]

[1]*Centre of Operations Research, Miguel Hernández University of Elche,* [a]*ecabello@umh.es,* [b]*d.morales@umh.es*
[2]*Department of Economic and Financial Studies, Miguel Hernández University of Elche,* [c]*agustin.perez@umh.es*

Exposure indices measure the degree of contact between two groups and are used to quantify occupational discrepancies between genders in a set of occupational sectors. This paper presents a novel methodology for predicting area-level proportions of employed men and women across various occupation sectors, along with estimating exposure indexes. The challenge arises from the compositional nature of the direct estimators of proportions, which tend to be imprecise when sample sizes are small. To overcome this problem, we propose to use a compositional multivariate Fay–Herriot model. By applying log-ratio transformations to the direct estimators of proportions, we can effectively capture the underlying structure and dependencies within the data. Small area estimators for proportions and exposure indexes are derived from the fitted model, and their corresponding root-mean-squared errors are estimated using parametric bootstrap techniques. To demonstrate the applicability of our approach, we conduct a case study using data from quarters 3 and 4 of the Spanish Labour Force Survey of 2022. The primary objective is to investigate the state of gender occupational segregation in Spanish provinces, thereby providing valuable insights into this socioeconomic phenomenon.

REFERENCES

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

ARIMA, S., BELL, W. R., DATTA, G. S., FRANCO, C. and LISEO, B. (2017). Multivariate Fay-Herriot Bayesian estimation of small area means under functional measurement error. *J. Roy. Statist. Soc. Ser. A* **180** 1191–1209. MR3723792 https://doi.org/10.1111/rssa.12321

BELL, W. (1954). A probability model for the measurement of ecological segregation. *Soc. Forces* **32** 357–364.

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

BENAVENT, R. and MORALES, D. (2021). Small area estimation under a temporal bivariate area-level linear mixed model with independent time effects. *Stat. Methods Appl.* **30** 195–222. MR4221680 https://doi.org/10.1007/s10260-020-00521-x

BERG, E. J. and FULLER, W. A. (2012). Estimators of error covariance matrices for small area prediction. *Comput. Statist. Data Anal.* **56** 2949–2962. MR2929350 https://doi.org/10.1016/j.csda.2012.02.030

BOUBETA, M., LOMBARDÍA, M. J. and MORALES, D. (2016). Empirical best prediction under area-level Poisson mixed models. *TEST* **25** 548–569. MR3531842 https://doi.org/10.1007/s11749-015-0469-8

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

BURGARD, J. P., ESTEBAN, M. D., MORALES, D. and PÉREZ, A. (2020). A Fay-Herriot model when auxiliary variables are measured with error. *TEST* **29** 166–195. MR4063387 https://doi.org/10.1007/s11749-019-00649-3

BURGARD, J. P., ESTEBAN, M. D., MORALES, D. and PÉREZ, A. (2021). Small area estimation under a measurement error bivariate Fay-Herriot model. *Stat. Methods Appl.* **30** 79–108. MR4221675 https://doi.org/10.1007/s10260-020-00515-9

Key words and phrases. Small area estimation, multivariate Fay–Herriot model, compositional data, bootstrap, exposure index, ocuppation sectors, labour force survey.

BURGARD, J. P., MORALES, D. and WÖLWER, A.-L. (2022). Small area estimation of socioeconomic indicators for sampled and unsampled domains. *AStA Adv. Stat. Anal.* **106** 287–314. MR4426858 https://doi.org/10.1007/s10182-021-00426-4

CABELLO, E., MORALES, D. and PÉREZ, A. (2025). Supplement to "Predicting gender employment discrepancies: A multivariate Fay-Herriot model for transformed proportions." https://doi.org/10.1214/25-AOAS2020SUPP

CHAMBERS, R., SALVATI, N. and TZAVIDIS, N. (2016). Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. *J. Roy. Statist. Soc. Ser. A* **179** 453–479. MR3461591 https://doi.org/10.1111/rssa.12123

DATTA, G. S., LAHIRI, P. and MAITI, T. (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *J. Statist. Plann. Inference* **102** 83–97. MR1885193 https://doi.org/10.1016/S0378-3758(01)00173-2

DATTA, G. S., LAHIRI, P., MAITI, T. and LU, K. L. (1999). Hierarchical Bayes estimation of unemployment rates for the U.S. states. *J. Amer. Statist. Assoc.* **94** 1074–1082.

EGOZCUE, J. J. and PAWLOWSKY-GLAHN, V. (2019). Compositional data: The sample space and its structure. *TEST* **28** 599–638. MR3992128 https://doi.org/10.1007/s11749-019-00670-6

ERCIULESCU, A. L. and OPSOMER, J. D. (2022). A model-based approach to predict employee compensation components. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **71** 1503–1520. MR4511120 https://doi.org/10.1111/rssc.12587

ESTEBAN, M. D., LOMBARDÍA, M. J., LÓPEZ-VIZCAÍNO, E., MORALES, D. and PÉREZ, A. (2020). Small area estimation of proportions under area-level compositional mixed models. *TEST* **29** 793–818. MR4140784 https://doi.org/10.1007/s11749-019-00688-w

ESTEBAN, M. D., MORALES, D., PÉREZ, A. and SANTAMARÍA, L. (2012). Small area estimation of poverty proportions under area-level time models. *Comput. Statist. Data Anal.* **56** 2840–2855. MR2929342 https://doi.org/10.1016/j.csda.2011.10.015

FABRIZI, E., FERRANTE, M. R. 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* (M. Pratesi, ed.) Wiley, New York.

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

FERRANTE, M. R. and TRIVISANO, C. (2010). Small area estimation of the number of firms' recruits by using multivariate models for count data. *Surv. Methodol.* **36** 171–180.

FRANCO, C. and BELL, W. R. (2014). Borrowing information over time in binomial/logit normal models for small area estimation. *Stat. Transit.* **16** 563–584.

GHOSH, M., NANGIA, N. and KIM, D. H. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *J. Amer. Statist. Assoc.* **91** 1423–1431. MR1439084 https://doi.org/10.2307/2291568

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D. and SANTAMARÍA, L. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Comput. Statist. Data Anal.* **52** 5242–5252. MR2526589 https://doi.org/10.1016/j.csda.2008.04.031

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D. and SANTAMARÍA, L. (2010). Small area estimation under Fay-Herriot models with non-parametric estimation of heteroscedasticity. *Stat. Model.* **10** 215–239. MR2759927 https://doi.org/10.1177/1471082X0801000206

GUHA, S. and CHANDRA, H. (2024). Small area estimation under a spatially correlated multivariate area-level model. *J. Roy. Statist. Soc. Ser. A* **187** 62–84. MR4754078 https://doi.org/10.1093/jrsssa/qnad079

HUANG, E. and BELL, W. (2004). An empirical study on using ACS supplementary survey data in SAIPE state poverty models. In 2004 *Proceedings of the American Statistical Association* 3677–3684. U.S. Bureau of the Census, Washington DC.

KRAUSE, J., BURGARD, J. P. and MORALES, D. (2022). ℓ_2-penalized approximate likelihood inference in logit mixed models for regional prevalence estimation under covariate rank-deficiency. *Metrika* **85** 459–489. MR4405684 https://doi.org/10.1007/s00184-021-00837-y

LÓPEZ-VIZCAÍNO, E., LOMBARDÍA, M. J. and MORALES, D. (2013). Multinomial-based small area estimation of labour force indicators. *Stat. Model.* **13** 153–178. MR3179522 https://doi.org/10.1177/1471082X13478873

LÓPEZ-VIZCAÍNO, E., LOMBARDÍA, M. J. and MORALES, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *J. Roy. Statist. Soc. Ser. A* **178** 535–565. MR3348348 https://doi.org/10.1111/rssa.12085

MARHUENDA, Y., MOLINA, I., MORALES, D. and RAO, J. N. K. (2017). Poverty mapping in small areas under a twofold nested error regression model. *J. Roy. Statist. Soc. Ser. A* **180** 1111–1136. MR3723789 https://doi.org/10.1111/rssa.12306

MARHUENDA, Y., MORALES, D. and PARDO, M. C. (2016). Tests for the variance parameter in the Fay-Herriot model. *Statistics* **50** 27–42. MR3435167 https://doi.org/10.1080/02331888.2015.1016026

MARHUENDA, Y., MORALES, D. and PARDO, M. D. C. (2014). Information criteria for Fay-Herriot model selection. *Comput*. *Statist*. *Data Anal*. **70** 268–280. MR3125493 https://doi.org/10.1016/j.csda.2013.09.016

MASSEY, D. S. and DENTON, N. A. (1988). The dimensions of residential segregation. *Soc*. *Forces* **67** 281–315. https://doi.org/10.1093/sf/67.2.281

MOLINA, I., SAEI, A. and LOMBARDÍA, M. J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *A* **170** 975–1000. MR2408988 https://doi.org/10.1111/j.1467-985X.2007.00493.x

MORALES, D., ESTEBAN, M. D., PÉREZ, A. and HOBZA, T. (2021). *A Course on Small Area Estimation and Mixed Models—Methods*, *Theory and Applications in R*. *Statistics for Social and Behavioral Sciences*. Springer, Cham. MR4244367 https://doi.org/10.1007/978-3-030-63757-6

MORALES, D., KRAUSE, J. and BURGARD, J. P. (2022). On the use of aggregate survey data for estimating regional major depressive disorder prevalence. *Psychometrika* **87** 344–368. MR4410428 https://doi.org/10.1007/s11336-021-09808-8

MORALES, D., PAGLIARELLA, M. C. and SALVATORE, R. (2015). Small area estimation of poverty indicators under partitioned area-level time models. *SORT* **39** 19–34. MR3381784

PAWLOWSKY-GLAHN, V. and BUCCIANTI, A. (2011). *Compositional Data Analysis*. Wiley, Chichester. https://doi.org/10.1002/9781119976462

PFEFFERMANN, D. and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Surv*. *Methodol*. **16** 217–237. https://doi.org/10.2307/3315407

PORTER, A. T., WIKLE, C. K. and HOLAN, S. H. (2015). Small area estimation via multivariate Fay-Herriot models with latent spatial dependence. *Aust*. *N*. *Z*. *J*. *Stat*. **57** 15–29. MR3335325 https://doi.org/10.1111/anzs.12101

PRATESI, M. (2016). *Analysis of Poverty Data by Small Area Estimation*. *Wiley Series in Survey Methodology*. Wiley, New York. https://doi.org/10.1002/9781118814963

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

RAO, J. N. K. and YU, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *Canad*. *J*. *Statist*. **22** 511–528. MR1321472 https://doi.org/10.2307/3315407

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

SINGH, B., SHUKLA, G. and KUNDU, D. (2005). Spatio-temporal models in small area estimation. *Surv*. *Methodol*. **31** 183–195.

SOUZA, D. B. and MOURA, F. A. S. (2016). Multivariate beta regression with applications in small area estimation. *J*. *Off*. *Stat*. **32** 747–768. https://doi.org/10.1515/jos-2016-0038

TZAVIDIS, N., RANALLI, M. G., SALVATI, N., DREASSI, E. and CHAMBERS, R. (2015). Robust small area prediction for counts. *Stat*. *Methods Med*. *Res*. **24** 373–395. MR3350290 https://doi.org/10.1177/0962280214520731

UBAIDILLAH, A., NOTODIPUTRO, K. A., KURNIA, A. and MANGKU, I. W. (2019). Multivariate Fay-Herriot models for small area estimation with application to household consumption per capita expenditure in Indonesia. *J*. *Appl*. *Stat*. **46** 2845–2861. MR4007698 https://doi.org/10.1080/02664763.2019.1615420

YOU, Y. and RAO, J. N. K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Surv*. *Methodol*. **26** 173–181.

ZHANG, L.-C. and CHAMBERS, R. L. (2004). Small area estimates for cross-classifications. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **66** 479–496. MR2062389 https://doi.org/10.1111/j.1369-7412.2004.05266.x

The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1778–1799
https://doi.org/10.1214/24-AOAS1992

COMPARING BASEBALL PLAYERS ACROSS ERAS VIA NOVEL FULL HOUSE MODELING

BY SHEN YAN[1,a], ADRIAN BURGOS, JR.[2,d], CHRISTOPHER KINSON[1,b] AND DANIEL J. ECK[1,c]

[1]*Department of Statistics, University of Illinois Urbana-Champaign,* [a]*shenyan3@illinois.edu,* [b]*kinson2@illinois.edu,* [c]*dje13@illinois.edu*

[2]*Department of History, University of Illinois Urbana-Champaign,* [d]*burgosjr@illinois.edu*

A new methodological framework suitable for era-adjusting baseball statistics is developed in this article. Within this methodological framework specific models are motivated. We call these models Full House Models. Full House Models work by balancing the achievements of Major League Baseball (MLB) players within a given season and the size of the MLB talent pool from which a player came. We demonstrate the utility of Full House Models in an application of comparing baseball players' performance statistics across eras. Our results reveal a new ranking of baseball's greatest players which include several modern players among the top all-time players. Modern players are elevated by Full House Modeling because they come from a larger talent pool. We present sensitivity and multiverse analyses to examine how changes in modeling inputs, including the estimate of the talent pool, affect the results.

REFERENCES

FRIENDLY, M., DALZELL, C., MONKMAN, M., MURPHY, D., FOOT, V., ZAKI-AZAT, J. and LAHMAN, S. (2023). Lahman: Sean 'Lahman' Baseball Database R package version 11.0-0.

ALBERT, J. and BENNETT, J. (2001). *Curve Ball*: *Baseball*, *Statistics*, *and the Role of Chance in the Game*. Copernicus, New York. MR1835464 https://doi.org/10.1007/b97222

ARMOUR, M. (2007). The effects of integration, 1947–1986. *Baseb*. *Res*. *J*. **36** 53–57.

ARMOUR, M. and LEVITT, D. R. (2016). Baseball demographics, 1947–2016. Society for American Baseball Research.

BAUMER, B. S., JENSEN, S. T. and MATTHEWS, G. J. (2015). openWAR: An open source system for evaluating overall player performance in major league baseball. *J. Quant*. *Anal*. *Sports* **11** 69–84.

BAUMER, B. S. and MATTHEWS, G. J. (2014). A statistician reads the sports page: There is no avoiding WAR. *Chance* **27** 41–44.

BELLAMY, R. V. and WALKER, J. R. (2004). Did televised baseball kill the "golden age" of the minor leagues? NINE: A. *J. Baseb*. *Hist*. *Cult*. **13** 59–73.

BERRI, D. and SCHMIDT, M. (2010). *Stumbling on Wins* (*Bonus Content Edition*): *Two Economists Expose the Pitfalls on the Road to Victory in Professional Sports*. *Portable Documents*. Pearson Education, Upper Saddle River.

BERRY, S. M., REESE, C. S. and LARKEY, P. D. (1999). Bridging different eras in sports. *J. Amer. Statist*. *Assoc*. **94** 661–676.

Baseball Reference Bullpen (2023). Wins Above Replacement.

BURGOS, A. JR (2007). *Playing America's Game*. Univ. California Press, Berkeley.

DAROWSKI, A., CHUPP, J. and BERKOWITZ, M. (2023). Hall of Stats (2012–2023).

ECK, D. J. (2020a). Challenging nostalgia and performance metrics in baseball. *Chance* **33** 16–25.

ECK, D. J. (2020b). Challenging WAR and other statistics as era-adjustment tools. FanGraphs. https://community.fangraphs.com/challenging-war-and-other-statistics-as-era-adjustment-tools/.

ESPN (2022). *Top* 100 *MLB Players of All Time*.

FAINARU-WADA, M. and QUINN, T. J. (2021). Barry Bonds, Roger Clemens are far less great if you subtract the PED factor. ESPN.

FORMAN, S. (2010). *Player Wins Above Replacement*.

Key words and phrases. Order statistics, nonparametric estimation, tail-probability modeling, era-adjusted baseball statistics.

GOULD, S. J. (1996). *Full House*: *The Spread of Excellence from Plato to Darwin*. Harvard Univ. Press, Cambridge.

HOAGLIN, D. C. et al. (1983). Letter values: A set of selected order statistics. *Underst*. *Robust Explor.Data Anal*. 33–57.

HOROWITZ, I. (1997). The increasing competitive balance in major league baseball. *Rev. Ind*. *Organ*. **12** 373–387.

KACZYNSKI, W., LEEMIS, L., LOEHR, N. and MCQUESTON, J. (2012). Nonparametric random variate generation using a piecewise-linear cumulative distribution function. *Comm*. *Statist*. *Simulation Comput*. **41** 449–468. MR2868998 https://doi.org/10.1080/03610918.2011.606947

LAND, K. C., DAVIS, W. R. and BLAU, J. R. (1994). Organizing the boys of summer: The evolution of US minor-league baseball, 1883-1990. *Am*. *J*. *Sociol*. **100** 781–813.

LEENAERTS, D. M. and BOKHOVEN, W. M. V. (1998). *Piecewise Linear Modeling and Analysis*. Kluwer Academic, Norwell.

LEWIS, M. (2003). *Moneyball*: *The Art of Winning an Unfair Game*. Norton, New York.

MCCRACKEN, V. (2001). Pitching and defense: How much control do hurlers have? *Baseb. Prospectus* **23**.

NEWMAN, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemp*. *Phys*. **46** 323–351.

PARETO, V. (1964). Cours d'économie politique: Nouvelle édition par G. H. Bousquet et G. Busino, Librairie Droz, Geneva 299–345.

POSNANSKI, J. (2021). *The Baseball* 100. Avid Reader Press.

Baseball Prospectus Staff (2013). WARP.

RADER, B. G. (2008). *Baseball*: *A History of America's Game*, 3rd ed. Univ. Illinois Press, Champaign.

RIESS, S. A. (1980). Professional baseball and social mobility. *J*. *Interdiscip*. *Hist*. **11** 235–250.

SCHELL, M. J. (1999). *Baseball's All-Time Best Hitters*. Princeton Univ. Press, Princeton.

SCHELL, M. J. (2005). *Baseball's All-Time Best Sluggers*. Princeton Univ. Press, Princeton.

SCHMIDT, M. B. and BERRI, D. J. (2005). Concentration of playing talent: Evolution in major league baseball. *J*. *Sports Econ*. **6** 412–419.

SCHOLZ, F. (1995). *Nonparametric Tail Extrapolation*. *White Paper from Boeing Information & Support Services*.

SHAPIRO, S. S. and WILK, M. B. (1965). An analysis of variance test for normality: Complete samples. *Biometrika* **52** 591–611. MR0205384 https://doi.org/10.1093/biomet/52.3-4.591

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR0848134 https://doi.org/10.1007/978-1-4899-3324-9

SLOWINSKI, P. (2010). What is WAR? FanGraphs.

STEEGEN, S., TUERLINCKX, F., GELMAN, A. and VANPAEMEL, W. (2016). Increasing transparency through a multiverse analysis. *Perspect*. *Psychol*. *Sci*. **11** 702–712. https://doi.org/10.1177/1745691616658637

STEIN, M. L. (2021). A parametric model for distributions with flexible behavior in both tails. *Environmetrics* **32** Paper No. e2658, 24. MR4229751 https://doi.org/10.1002/env.2658

SULLIVAN, N. J. (1990). *The Minors*. Macmillan.

TANGO, T. M., LICHTMAN, M. G. and DOLPHIN, A. E. (2007). *The Book*: *Playing the Percentages in Baseball*. Potomac Books, Inc.

THORN, J. and PALMER, P. (1984). *The Hidden Game of Baseball*: *A Revolutionary Approach to Baseball and Its Statistics*. Univ. Chicago Press, Chicago.

YAN, S., BURGOS, JR., A., KINSON, C. and ECK, D. J (2025). Supplement to "Comparing baseball players across eras via novel Full House Modeling." https://doi.org/10.1214/24-AOAS1992SUPPA, https://doi.org/10.1214/24-AOAS1992SUPPB