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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles

Causal Inference, Design and Policy

Computational Biology

Environmental Science

continued

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

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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

continued

Continued on inner cover

Vol. 20, No. 2—June 2026

THE ANNALS OF APPLIED STATISTICS Vol. 20, No. 2, pp. 893–1787 June 2026

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

IMS OFFICERS

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

President-Elect: Richard J. Samworth, Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, CB3 0WB, UK

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

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

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

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

IMS PUBLICATIONS

The Annals of Statistics. *Editors*: Hans-Georg Müller, Department of Statistics, University of California, Davis, Davis, CA 95616, USA. Harrison Zhou, Department of Statistics and Data Science, Yale University, New Haven, CT 06520, USA

The Annals of Applied Statistics. *Editor-in-Chief*: Lexin Li, Department of Biostatistics and Epidemiology, University of California, Berkeley, Berkeley, CA 94720-7360, USA

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

The Annals of Applied Probability. *Editors*: Jian Ding, School of Mathematical Sciences, Peking University, 100871, Bejing, China. Claudio Landim, IMPA, 22461-320, Rio de Janeiro, Brazil

Statistical Science. *Editor*: Lutz Dümbgen, Institute of Mathematical Statistics and Actuarial Science, University of Bern, Alpeneggstrasse 22, CH-3012 Bern, Switzerland

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

The Annals of Applied Statistics **[ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 20, Number 2, June 2026. 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
2026, Vol. 20, No. 2, 893–920
https://doi.org/10.1214/26-AOAS2168

AUTOREGRESSIVE MODELS FOR PANEL DATA CAUSAL INFERENCE WITH APPLICATION TO STATE-LEVEL OPIOID POLICIES

BY JOSEPH ANTONELLI[1,a], MAX RUBINSTEIN[2,b], DENIS AGNIEL[2,c], ROSANNA SMART[2,d], ELIZABETH A. STUART[3,i], MATTHEW CEFALU[4,j], TERRY SCHELL[2,e], JOSHUA EAGAN[2,f], ELIZABETH STONE[5,k], MAX GRISWOLD[2,g] AND BETH ANN GRIFFIN[2,h]

[1]*Department of Statistics, University of Florida,* [a]*jantonelli@ufl.edu*

[2]*RAND Corporation,* [b]*mrubinstein@rand.org,* [c]*dagniel@rand.org,* [d]*rsmart@rand.org,* [e]*tschell@rand.org,* [f]*jeagan@rand.org,* [g]*griswold@rand.org,* [h]*bethg@rand.org*

[3]*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health,* [i]*estuart@jhu.edu*

[4]*Disney Streaming,* [j]*m.s.cefalu@gmail.com*

[5]*Department of Psychiatry, Rutgers Robert Wood Johnson Medical School,* [k]*elizabeth.stone@rutgers.edu*

Motivated by the study of state opioid policies, we propose a novel approach using autoregressive models for causal effect estimation in panel data settings. We estimate the impact of key opioid-related policies, specifically must-access prescription drug monitoring programs (PDMPs), naloxone access laws (NALs), and medical marijuana laws, on opioid prescribing. Existing methods, such as difference-in-differences and synthetic controls, are difficult to apply in dynamic policy environments with multiple overlapping policies and small sample sizes. While autoregressive models have been used in similar contexts, they have lacked formal justification until now. We outline assumptions that link these models to causal effects and study the bias of resulting estimates when key assumptions are violated. Through simulation studies mirroring our application, we show that our proposed estimators often outperform existing methods. We thus provide a formal justification for using autoregressive models to evaluate the effectiveness of state policies in addressing the opioid crisis.

REFERENCES

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

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

ACHEN, C. H. (2000). Why lagged dependent variables can suppress the explanatory power of other independent variables. In annual meeting of the political methodology section of the American political science association. *UCLA* **20** 7–2000.

AGARWAL, A., SHAH, D. and SHEN, D. (2020). Synthetic interventions. arXiv preprint. Available at arXiv:2006.07691.

ALPERT, A., POWELL, D. and PACULA, R. L. (2018). Supply-side drug policy in the presence of substitutes: Evidence from the introduction of abuse-deterrent opioids. *Amer. Econ. J., Econ. Policy* **10** 1–35.

AMERICAN COLLEGE OF EMERGENCY PHYSICIANS (2021). Opioid Regulations: State by State Guide. https://www.acep.org/siteassets/sites/acep/media/by-medical-focus/opioids/opioid-guide-state-by-state.pdf. Accessed September 14, 2022.

ANDRAKA-CHRISTOU, B., RANDALL-KOSICH, O., GOLAN, M., TOTARAM, R., SALONER, B., GORDON, A. J. and STEIN, B. D. (2022a). A national survey of state laws regarding medications for opioid use disorder in problem-solving courts. *Health Justice* **10** 14.

ANDRAKA-CHRISTOU, B., SALONER, B., GORDON, A. J., TOTARAM, R., RANDALL-KOSICH, O., GOLAN, M. and STEIN, B. D. (2022b). Laws for expanding access to medications for opioid use disorder: A legal analysis of 16 states & Washington D.C. *Amer. J. Drug Alcohol Abuse* **48** 492–503.

Key words and phrases. Autoregressive models, causal inference, panel data.

ANGRIST, J. D. and PISCHKE, J.-S. (2008). *Mostly Harmless Econometrics*: *An Empiricist's Companion*. Princeton University Press, Princeton.

ANTONELLI, J., RUBINSTEIN, M., AGNIEL, D., SMART, R., STUART, E. A., CEFALU, M., SCHELL, T., EAGAN, J., STONE, E., GRISWOLD, M. and GRIFFIN, B. A. (2026). Supplement to "Autoregressive models for panel data causal inference with application to state-level opioid policies." https://doi.org/10.1214/26-AOAS2168SUPP

ASHENFELTER, O. C. and CARD, D. (1984). Using the longitudinal structure of earnings to estimate the effect of training programs.

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

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

BENJAMIN, M. A., RIGBY, R. A. and STASINOPOULOS, D. M. (2003). Generalized autoregressive moving average models. *J. Amer. Statist. Assoc.* **98** 214–223. MR1965687 https://doi.org/10.1198/016214503388619238

BUCHMUELLER, T. C. and CAREY, C. (2018). The effect of prescription drug monitoring programs on opioid utilization in Medicare. *Amer. Econ. J., Econ. Policy* **10** 77–112.

BURRIS, S., JOHNSON, K., IBRAHIM, J., PLATT, E. and ALLEN, L. (2017). State-level opioid antagonist access laws: The emergence of three distinct strategies, 2001-2015. *Drug Alcohol Depend.* **191** e29.

CALLAWAY, B. and SANT'ANNA, P. H. C. (2021). Difference-in-differences with multiple time periods. *Rev. Econ. Stat.* **103** 133–146. https://doi.org/10.1162/rest_a_01015

CARROLL, R. J. (1998). Measurement error in epidemiologic studies. *Encycl. Biostat.* **3** 2491–2519.

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

CAWLEY, J. and DRAGONE, D. (2023). Harm reduction: When does it improve health, and when does it backfire? Working Paper No. 30926, National Bureau of Economic Research.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2020a). Wide-ranging Online Data for Epidemiologic Research (WONDER). https://wonder.cdc.gov/. Accessed: 2023-08-23.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2020b). Overdose Deaths Accelerating During COVID-19. https://www.cdc.gov/media/releases/2020/p1218-overdose-deaths-covid-19.html. Accessed: 2023-07-20.

CHAMBERLAIN, J. A. and KLEIN, B. L. (1994). A comprehensive review of naloxone for the emergency physician. *Amer. J. Emerg. Med.* **12** 650–660.

DAVIS, C. and CARR, D. (2017). State legal innovations to encourage naloxone dispensing. *J. Amer. Pharm. Assoc.* **57** S180–S184.

DAVIS, C. S. and CARR, D. (2015). Legal changes to increase access to naloxone for opioid overdose reversal in the United States. *Drug Alcohol Depend.* **157** 112–120.

DAVIS, C. S. and LIEBERMAN, A. J. (2021). Laws limiting prescribing and dispensing of opioids in the United States, 1989-2019. *Addiction* **116** 1817–1827.

DAVIS, C. S. and SAMUELS, E. A. (2020). Opioid policy changes during the COVID-19 pandemic—and beyond. *J. Addict. Med.* **14** e4–e5.

DAVIS, C. S. and SAMUELS, E. A. (2021). Continuing increased access to buprenorphine in the United States via telemedicine after COVID-19. *Int. J. Drug Policy* **93** 102905.

ELWERT, F. and WINSHIP, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annu. Rev. Sociol.* **40** 31–53. https://doi.org/10.1146/annurev-soc-071913-043455

EVANS, W. N., LIEBER, E. M. and POWER, P. (2019). How the reformulation of OxyContin ignited the heroin epidemic. *Rev. Econ. Stat.* **101** 1–15.

FRANK, R., HUMPHREYS, K. and POLLACK, H. (2021). Policy responses to the addiction crisis. *J. Health Polit. Policy Law* **46** 585–597.

FRICKE, H. (2017). Identification based on difference-in-differences approaches with multiple treatments. *Oxf. Bull. Econ. Stat.* **79** 426–433.

GILL, R. D. and ROBINS, J. M. (2001). Causal inference for complex longitudinal data: The continuous case. *Ann. Statist.* **29** 1785–1811. MR1891746 https://doi.org/10.1214/aos/1015345962

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

GREEN, T. C., DAVIS, C., XUAN, Z., WALLEY, A. Y. and BRATBERG, J. (2020). Laws mandating coprescription of naloxone and their impact on naloxone prescription in five US states, 2014–2018. *Amer. J. Publ. Health* **110** 881–887.

GRIFFIN, B. A., SCHULER, M., STONE, E. M., PATRICK, S. W., STEIN, B. D., SCHERLING, A. and STUART, E. A. (2023). Identifying optimal methods for addressing confounding bias when estimating the effects of state-level policies. *Epidemiology* **34** 856–864. https://doi.org/10.1097/EDE.0000000000001659

GRIFFIN, B. A., SCHULER, M. S., PANE, J., PATRICK, S. W., SMART, R., BRADLEY, D. S., GRIMM, G. and STUART, E. (2022). Methodological considerations for estimating policy effects in the context of co-occurring policies. *Health Serv. Outcomes Res. Methodol.* https://doi.org/10.1007/s10742-022-00284-w

GRIFFIN, B. A., SCHULER, M. S., STUART, E. A., PATRICK, S., MCNEER, E., SMART, R., POWELL, D., STEIN, B. D., SCHELL, T. L. et al. (2021). Moving beyond the classic difference-in-differences model: A simulation study comparing statistical methods for estimating effectiveness of state-level policies. *BMC Med. Res. Methodol.* **21** 279. https://doi.org/10.1186/s12874-021-01471-y

GRISWOLD, M., GRIFFIN, B. A., SCHULER, M., RUBENSTEIN, M., STONE, E., STEIN, B. and STUART, E. (2023). Simulating Time-Varying Estimators to Understand Potential Bias and Precision in Policy Effects. Unpublished.

HAFFAJEE, R. L., BOHNERT, A. S. B. and LAGISETTY, P. A. (2018). Policy pathways to address provider workforce barriers to buprenorphine treatment. *Amer. J. Prev. Med.* **54** S230–S242.

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

HORWITZ, J., DAVIS, C. S., MCCLELLAND, L. S., FORDON, R. S. and MEARA, E. (2018). The problem of data quality in analyses of opioid regulation: The case of prescription drug monitoring programs. Working Paper No. 24947, National Bureau of Economic Research. https://doi.org/10.3386/w24947

IMAI, K. and KIM, I. S. (2021). On the use of two-way fixed effects regression models for causal inference with panel data. *Polit. Anal.* **29** 405–415.

MILLS, H. L., HIGGINS, J. P., MORRIS, R. W., KESSLER, D., HERON, J., WILES, N., SMITH, G. D. and TILLING, K. (2021). Detecting heterogeneity of intervention effects using analysis and meta-analysis of differences in variance between trial arms. *Epidemiology* **32** 846–854.

NGUYEN, T. Q., SCHMID, I. and STUART, E. A. (2021). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychol. Methods* **26** 255.

NICKELL, S. (1981). Biases in dynamic models with fixed effects. *Econometrica* **49** 1417–1426. MR0636160 https://doi.org/10.2307/1911408

PACULA, R. L., POWELL, D., HEATON, P. and SEVIGNY, E. (2015). Assessing the effects of medical marijuana laws on marijuana: The devil is in the details. *J. Policy Anal. Manage.* **34** 7–31.

PACULA, R. L. and STEIN, B. D. (2020). State Approaches to Tackling the Opioid Crisis Though the Health Care System. Brookings Report Published as Part of the Opioid Crisis in America: Domestic and International Dimensions.

PATEL, I., WALTER, L. A. and LI, L. (2021). Opioid overdose crises during the COVID-19 pandemic: Implication of health disparities. *Harm Reduct. J.* **18** 89.

PATRICK, S. W., FRY, C. E., JONES, T. F. and BUNTIN, M. B. (2016). Implementation of prescription drug monitoring programs associated with reductions in opioid-related death rates. *Health Aff.* **35** 1324–1332.

PEARL, J. (2009). *Causality*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

POWELL, D., PACULA, R. L. and JACOBSON, M. (2018). Do medical marijuana laws reduce addictions and deaths related to pain killers? *J. Health Econ.* **58** 29–42. https://doi.org/10.1016/j.jhealeco.2017.12.007

RAND-USC SCHAEFFER OPIOID POLICY TOOLS AND INFORMATION CENTER (2024a). OPTIC-Vetted Medical Marijuana Policy Data. Obtained from https://www.rand.org/health-care/centers/optic/resources/datasets.html on January 15, 2024.

RAND-USC SCHAEFFER OPIOID POLICY TOOLS AND INFORMATION CENTER (2024c). OPTIC-Vetted Naloxone Policy Data. Obtained from https://www.rand.org/health-care/centers/optic/resources/datasets.html on January 15, 2024.

RAND-USC SCHAEFFER OPIOID POLICY TOOLS AND INFORMATION CENTER (2024b). Opioid Policy Tools and Information Center (2024b). OPTIC-Vetted PDMP Policy Data. Obtained from https://www.rand.org/health-care/centers/optic/resources/datasets.html on January 15, 2024.

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

ROLLER, M. and STEINBERG, D. (2023). *Differences-in-Differences with Multiple Treatments Under Control.*

SABATIER, P. and MAZMANIAN, D. (1980). The implementation of public policy: A framework of analysis. *Policy Stud. J.* **8** 538–560.

SCHELL, T. L., GRIFFIN, B. A. and MORRAL, A. R. (2018). *Evaluating Methods to Estimate the Effect of State Laws on Firearm Deaths*: *A Simulation Study*. RAND Corporation.

SCHULER, M. S., HEINS, S. E., SMART, R., GRIFFIN, B. A., POWELL, D., STUART, E. A., PARDO, B., SMUCKER, S., PATRICK, S. W. et al. (2020). The state of the science in opioid policy research. *Drug Alcohol Depend.* **214** 108137. https://doi.org/10.1016/j.drugalcdep.2020.108137

SMART, R. and GRANT, S. (2021). Effectiveness and implementability of state-level naloxone access policies: Expert consensus from an online modified-delphi process. *Int. J. Drug Policy* **98** 103383.

SMART, R., HAFFAJEE, R. L. and DAVIS, C. S. (2022). Legal review of state emergency medical services policies and protocols for naloxone administration. *Drug Alcohol Depend*. **238** 109589.

SMART, R., PARDO, B. and DAVIS, C. S. (2021). Systematic review of the emerging literature on the effectiveness of naloxone access laws in the United States. *Addiction* **116** 6–17.

SMART, R., POWELL, D., PACULA, R. L., PEET, E. D., ABOUK, R. and DAVIS, C. S. (2023). Investigating the complexity of naloxone distribution: Which policies matter for pharmacies and potential recipients. Working Paper No. 31142, National Bureau of Economic Research. https://doi.org/10.3386/w31142

VOLKOW, N. D. and COLLINS, F. S. (2017). The role of science in addressing the opioid crisis. *N. Engl. J. Med.* **377** 391–394.

WEN, H., HOCKENBERRY, J. M. and CUMMINGS, J. R. (2015). The effect of medical marijuana laws on adolescent and adult use of marijuana, alcohol, and other substances. *J. Health Econ*. **42** 64–80.

WOOLDRIDGE, J. M. (2023). Simple approaches to nonlinear difference-in-differences with panel data. *Econom. J.* **26** C31–C66. MR4643836 https://doi.org/10.1093/ectj/utad016

ZEGER, S. L. and QAQISH, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics* **44** 1019–1031. MR0980997 https://doi.org/10.2307/2531732

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 921–944
https://doi.org/10.1214/26-AOAS2155

MODEL-BASED SPATIAL RECONSTRUCTION OF LARGE-SCALE BIOMOLECULES VIA BAYESIAN INFERENCE OF A HIERARCHICAL SPATIAL MODEL

BY YINGCHENG LUO[1,a], SHIYU WANG[2,e], CHONG SHEN[1,b], YICHAO LI[1,c], ZHAOHUI S. QIN[2,f] AND KE DENG[1,d]

[1]*Department of Statistics and Data Science, Tsinghua University,* [a]*lyc22@mails.tsinghua.edu.cn,* [b]*c-shen18@mails.tsinghua.edu.cn,* [c]*ycli@mail.tsinghua.edu.cn,* [d]*kdeng@tsinghua.edu.cn*

[2]*Department of Biostatistics and Bioinformatics, Emory University,* [e]*shiyu.wang@emory.edu,* [f]*zhaohui.qin@emory.edu*

Revealing the spatial organization of biomolecules and characterizing their spatial distribution in cells and tissues have long been recognized as important problems in biomedical research. With rapid advances in DNA sequencing technologies in recent years, creative sequencing-based experimental assays, for example, Hi-C and DNA microscopy, have been invented to reveal the spatial properties of large-scale biomolecules in a high-throughput and high-resolution manner. A typical experiment based on these technologies produces a count matrix to record the contact frequencies among molecules of interest, which are closely associated with their spatial distances, allowing us to reconstruct the spatial organization of large-scale biomolecules via data analysis. There is a great appeal to develop statistically rigorous and computationally scalable methods for this important problem. In this study, we fill this gap with a novel method named HiSpa. Equipped with a hierarchical spatial model, HiSpa utilizes the idea of multiscale modeling to reduce the computational complexity from $O(n^2)$ to $O(n^2/T)$ with minimal loss of reconstruction quality, where T typically increases with n sublinearly. Advanced Monte Carlo strategies are developed for efficient Bayesian inference of HiSpa. The superiority of HiSpa over existing methods is demonstrated by simulation studies and real data applications.

REFERENCES

ABBAS, A., HE, X. et al. (2019). Integrating Hi-C and FISH data for modeling of the 3D organization of chromosomes. *Nat. Commun.* **10** 1–14.

ALBERTS, B., JOHNSON, A. et al. (2002). Genesis, modulation, and regeneration of skeletal muscle. In *Molecular Biology of the Cell*, 4th ed. Garland Science.

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

BINGHAM, G. C., LEE, F. et al. (2020). Spatial-omics: Novel approaches to probe cell heterogeneity and extracellular matrix biology. *Matrix Biol.* **91** 152–166.

CAO, Z., ZUO, W., WANG, L., CHEN, J., QU, Z., JIN, F. and DAI, L. (2023). Spatial profiling of microbial communities by sequential FISH with error-robust encoding. *Nat. Commun.* **14** 1477.

DIXON, J. R., SELVARAJ, S. et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485** 376–380.

DUAN, Z., ANDRONESCU, M. et al. (2010). A three-dimensional model of the yeast genome. *Nature* **465** 363–367.

FU, J., LIU, M. et al. (2012). Spatially-interactive biomolecular networks organized by nucleic acid nanostructures. *Acc. Chem. Res.* **45** 1215–1226.

HOVENGA, V., KALITA, J. and OLUWADARE, O. (2023). HiC-GNN: A generalizable model for 3D chromosome reconstruction using graph convolutional neural networks. *Comput. Struct. Biotechnol. J.* **21** 812–836.

HU, M., DENG, K. et al. (2012). HiCNorm: Removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28** 3131–3133.

Key words and phrases. Optics-free biological imaging, DNA microscopy, Hi-C, inverse problem, multiscale modelling.

HU, M., DENG, K. et al. (2013). Bayesian inference of spatial organizations of chromosomes. *PLoS Comput. Biol.* **9** e1002893.

JI, X. and ZHA, H. (2004). Sensor positioning in wireless ad-hoc sensor networks using multidimensional scaling. In *IEEE INFOCOM* 2004 **4** 2652–2661. IEEE Press, New York.

KALHOR, R., TJONG, H. et al. (2011). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30** 90–98.

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

LI, Z., PORTILLO-LEDESMA, S. and SCHLICK, T. (2023). Techniques for and challenges in reconstructing 3D genome structures from 2D chromosome conformation capture data. *Curr. Opin. Cell Biol.* **83** 102209.

LIEBERMAN-AIDEN, E., VAN BERKUM, N. L. et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326** 289–293.

LUO, Y., WANG, S., SHEN, C., LI, Y., QIN, Z. S. and DENG, K. (2026). Supplement to "Model-based spatial reconstruction of large-dcale biomolecules via Bayesian inference of a hierarchical spatial model." https://doi.org/10.1214/26-AOAS2155SUPPA, https://doi.org/10.1214/26-AOAS2155SUPPB

MAO, G., FIDAN, B. and ANDERSON, B. D. (2007). Wireless sensor network localization techniques. *Comput. Netw.* **51** 2529–2553.

NAGANO, T., LUBLING, Y. et al. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502** 59–64.

NIU, T., QIN, Z. S. et al. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Amer. J. Hum. Genet.* **70** 157–169.

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

OLUWADARE, O., ZHANG, Y. and CHENG, J. (2018). A maximum likelihood algorithm for reconstructing 3D structures of human chromosomes from chromosomal contact data. *BMC Genomics* **19** 161.

PINKEL, D., STRAUME, T. and GRAY, J. W. (1986). Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proc. Natl. Acad. Sci. USA* **83** 2934–2938.

QIN, Z. S., NIU, T. and LIU, J. S. (2002). Partition-ligation–expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Amer. J. Hum. Genet.* **71** 1242–1247.

RAO, S. S., HUNTLEY, M. H. et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159** 1665–1680.

ROUSSEAU, M., FRASER, J. et al. (2011). Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinform.* **12** 1–16.

SANG, H. and HUANG, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 111–132. MR2885842 https://doi.org/10.1111/j.1467-9868.2011.01007.x

SERRA, F., BAÙ, D. et al. (2017). Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput. Biol.* **13** e1005665.

SHAH, S., LUBECK, E., ZHOU, W. and CAI, L. (2016). In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92** 342–357.

SINGER, A. (2008). A remark on global positioning from local distances. *Proc. Natl. Acad. Sci. USA* **105** 9507–9511. MR2430205 https://doi.org/10.1073/pnas.0709842104

SOLOVEI, I., CAVALLO, A., SCHERMELLEH, L., JAUNIN, F., SCASSELATI, C., CMARKO, D., CREMER, C., FAKAN, S. and CREMER, T. (2002). Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence *in situ* hybridization (3D-FISH). *Exp. Cell Res.* **276** 10–23.

STEVENS, T. J., LANDO, D. et al. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544** 59–64.

SU, J.-H., ZHENG, P. et al. (2020). Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell* **182** 1641–1659.

TAN, L., XING, D. et al. (2018). Three-dimensional genome structures of single diploid human cells. *Science* **361** 924–928.

TRIEU, T., OLUWADARE, O. and CHENG, J. (2019). Hierarchical reconstruction of high-resolution 3D models of large chromosomes. *Sci. Rep.* **9** 1–12.

VAN BERKUM, N. L., LIEBERMAN-AIDEN, E. et al. (2010). Hi-C: A method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **39** e1869.

WANG, H., YANG, J., ZHANG, Y., QIAN, J. and WANG, J. (2022). Reconstruct high-resolution 3D genome structures for diverse cell-types using FLAMINGO. *Nat. Commun.* **13** 2645.

WANG, Y., EDDISON, M., FLEISHMAN, G., WEIGERT, M., XU, S., WANG, T., ROKICKI, K., GOINA, C., HENRY, F. E. et al. (2021). EASI-FISH for thick tissue defines lateral hypothalamus spatio-molecular organization. *Cell* **184** 6361–6377.

WEINSTEIN, J. A., REGEV, A. and ZHANG, F. (2019). DNA microscopy: Optics-free spatio-genetic imaging by a stand-alone chemical reaction. *Cell* **178** 229–241.

XIA, C., FAN, J. et al. (2019). Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci. USA* **116** 19490–19499.

XIAO, G., WANG, X. and KHODURSKY, A. B. (2011). Modeling three-dimensional chromosome structures using gene expression data. *J. Amer. Statist. Assoc.* **106** 61–72. MR2816702 https://doi.org/10.1198/jasa.2010.ap09504

YAFFE, E. and TANAY, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* **43** 1059–1065.

ZHANG, M., EICHHORN, S. W., ZINGG, B. et al. (2021). Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* **598** 137–143.

ZHANG, R., HU, M. et al. (2020). Inferring spatial organization of individual topologically associated domains via piecewise helical model. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17** 647–656.

ZHOU, F., HE, K., LI, Q., CHAPKIN, R. S. and NI, Y. (2022). Bayesian biclustering for microbial metagenomic sequencing data via multinomial matrix factorization. *Biostatistics* **23** 891–909. MR4454961 https://doi.org/10.1093/biostatistics/kxab002

ZOU, C., ZHANG, Y. and OUYANG, Z. (2016). HSA: Integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. *Genome Biol.* **17** 1–14.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 945–962
https://doi.org/10.1214/26-AOAS2164

MEASUREMENT ERROR MODELS FOR MORPHOMETRIC DATA

BY BEN C. STEVENSON[1,a], ELIZABETH SMIT[1,b] AND EDY SETYAWAN[2,c]

[1]*Department of Statistics, University of Auckland,* [a]*ben.stevenson@auckland.ac.nz,* [b]*esmi468@aucklanduni.ac.nz*
[2]*Elasmobranch Institute Indonesia,* [c]*edysetyawan@gmail.com*

An understanding of the body size of individuals and the relationships between different dimensions is critical for monitoring the status and the health of a wildlife population. Morphometric data have traditionally been collected by physically handling and measuring individual animals, but recent technological advancements allow researchers to deploy sophisticated but affordable instruments, like drones and camera traps, to take photos of individual animals from which morphometric measurements are extracted. However, morphometric data obtained from photographs can be less accurate than those from direct measurement. In this paper we propose a new linear mixed-effects model approach, involving multivariate normal distributions, to describe the latent true measurements and to accommodate measurement error. Our method can be used to estimate relationships between dimensions, to predict the measurement of any one dimension from any subset of other dimensions, and to make inference on allometry, including testing for allometric growth. We demonstrate the use of our method with an application to morphometric data of the reef manta ray *Mobula alfredi* collected by drones.

REFERENCES

AKRITAS, M. G. and BERSHADY, M. A. (1996). Linear regression for astronomical data with measurement errors and intrinsic scatter. *Astrophys. J.* **470** 706–714.

ARNOLD, T. W. and GREEN, A. J. (2007). On the allometric relationship between size and composition of avian eggs: A reassessment. *Ornithol. Appl.* **109** 705–714.

BARUZZI, C., SNOW, N. P., VERCAUTEREN, K. C., STRICKLAND, B. K., ARNOULT, J. S., FISCHER, J. W., GLOW, M. P., LAVELLE, M. J., SMITH, B. A. et al. (2023). Estimating body mass of wild pigs (*Sus scrofa*) using body morphometrics. *Ecol. Evol.* **13** e9853.

BIERLICH, K. C., SCHICK, R. S., HEWITT, J., DALE, J., GOLDBOGEN, J. A., FRIEDLAENDER, A. S. and JOHNSTON, D. W. (2021). Bayesian approach for predicting photogrammetric uncertainty in morphometric measurements derived from drones. *Mar. Ecol. Prog. Ser.* **673** 193–210.

BROWN, D. and COX, A. J. (2009). Innovative uses of video analysis. *Phys. Teach.* **47** 145–150.

CUI, S., CHEN, D., SUN, J., CHU, H., LI, C. and JIANG, Z. (2020). A simple use of camera traps for photogrammetric estimation of wild animal traits. *J. Zool.* **312** 12–20.

EATON, M. L. (1983). *Multivariate Statistics*: *A Vector Space Approach*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. MR0716321

FRANCIS, M. P. (2006). Morphometric minefields—towards a measurement standard for chondrichthyan fishes. *Environ. Biol. Fisches* **77** 407–421.

FRUCIANO, C. (2016). Measurement error in geometric morphometrics. *Dev. Genes Evol.* **226** 139–158.

FULLER, W. A. (1987). *Measurement Error Models*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. MR0898653 https://doi.org/10.1002/9780470316665

GAYON, J. (2000). History of the concept of allometry. *Amer. Zool.* **40** 748–758.

HODGSON, J. C., HOLMAN, D., TERAUDS, A., KOH, L. P. and GOLDSWORTHY, S. D. (2020). Rapid condition monitoring of an endangered marine vertebrate using precise, non-invasive morphometrics. *Biol. Conserv.* **242** 108402.

JIANG, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. *Springer Series in Statistics*. Springer, New York. MR2308058

KILMER, J. T. and RODRIGUEZ, R. L. (2016). Ordinary least squares regression is indicated for studies of allometry. *J. Evol. Biol.* **30** 4–12.

Key words and phrases. Allometry, drones, linear mixed-effects models, reef manta ray *Mobula alfredi*.

LAST, P., NAYLOR, G., SÉRET, B., WHITE, W., STEHMANN, M. and DE CARVALHO, M., eds. (2016). *Rays of the World*. CSIRO Publishing, Melbourne.

LINDSTROM, M. J. and BATES, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Amer. Statist. Assoc.* **83** 1014–1022. MR0997577

MCARDLE, B. H. (1988). The structural relationship: Regression in biology. *Can. J. Zool.* **66** 2329–2339.

MUIR, A. M., VECSEI, P. and KRUEGER, C. C. (2012). A perspective on perspectives: Methods to reduce variation in shape analysis of digital images. *Trans. Amer. Fish. Soc.* **141** 1161–1170.

NOTARBARTOLO-DI-SCIARA, G. (1987). A revisionary study of the genus *Mobula* Rafinesque, 1810 (chondrichtyes: Mobulidae) with the description of a new species. *Zool. J. Linn. Soc.* **91** 1–91.

PINHEIRO, J. C. and BATES, D. M. (1996). Unconstrained parameteriizations for variance-covariance matrices. *Stat. Comput.* **6** 289–296.

PINHEIRO, J. C. and BATES, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

PINHEIRO, J. C. and BATES, D. M. (2023). nlme: Linear and nonlinear mixed effects models. R package version 3.1-164.

SCHMIDT-NIELSEN, K. (1984). *Scaling*: *Why Is Animal Size so Important?* Cambridge Univ. Press, Cambridge.

SETYAWAN, E., STEVENSON, B. C., IZUAN, M., CONSTANTINE, R. and ERDMANN, M. V. (2022). How big is that manta ray? A novel and non-invasive method for measuring reef manta rays using small drones. *Drones* **6** 63.

SMITH, R. J. (2009). Use and misuse of the reduced major axis for line-fitting. *Amer. J. Phys. Anthropol.* **140** 476–486.

SOKAL, R. R. and ROHLF, F. J. (2012). *Biometry*. Freeman, New York.

STEVENSON, B. C., SMIT, E. and SETYAWAN, E. (2026). Supplement to "Measurement error models for morphometric data." https://doi.org/10.1214/26-AOAS2164SUPP

STEVENSON, B. C. and SMIT, E. (2026). morphErr: Linear mixed-effects models for morphometric data subject to measurement error. R package version 1.0.0. https://github.com/elismit01/morphErr.

TARUGARA, A., CLEGG, B. W., GANDIWA, E., MUPOSHI, V. K. and WENHAM, C. M. (2019). Measuring body dimensions of leopards (*Panthera pardus*) from camera trap photographs. *PeerJ* **7** e7630.

WARTON, D. I., WRIGHT, I. J., FALSTER, D. S. and WESTOBY, M. (2006). Bivariate line-fitting methods for allometry. *Biol. Rev.* **81** 259–291.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 963–985
https://doi.org/10.1214/26-AOAS2170

DECODING MICROBIOME DUAL MEDIATION: INTRODUCING ZIMMA FOR ENHANCED ZERO-INFLATED DATA ANALYSIS

BY XI QIAO[1,a], RUITAO LIU[2,b], XUEYING TANG[3,d], CHRISTINE B. PETERSON[4,e], ROBERT R. JENQ[5,f] AND LIANGLIANG ZHANG[2,c]

[1]*Epidemiology, Huntsman Cancer Institute, University of Utah,* [a]*xi.qiao@hci.utah.edu*
[2]*Population and Quantitative Health Sciences, Case Western Reserve University,* [b]*rxl761@case.edu,* [c]*lxz716@case.edu*
[3]*Mathematics, University of Arizona,* [d]*xytang@arizona.edu*
[4]*Biostatistics, The University of Texas MD Anderson Cancer Center,* [e]*CBPeterson@mdanderson.org*
[5]*Hematology & Hematopoietic Cell Transplantation, City of Hope,* [f]*rjenq@coh.org*

Given the complex interactions between the microbiome, the host, and external factors, causal mediation analysis is essential for unraveling how dysbiosis or microbial imbalance mediates the effects of interventions or environmental exposures on health outcomes. However, zero inflation in microbiome count data complicates high-dimensional mediation analysis, as frequently employed zero-inflated models often struggle to distinguish true zero inflation from the underlying count distribution. To address this, we employ a zero-inflated negative binomial distribution for each mediator within a Bayesian framework, incorporating spike-and-slab priors to enable sparsity in estimating natural indirect effects (NIE) and an informative prior based on nonzero counts to improve dispersion estimation and account for zero sources. Recognizing the distinct biological mechanisms underlying microbial presence vs. abundance, we developed a dual mediation model, ZIMMA, to separate the NIE into pathways for mediator abundance and prevalence. Extensive simulations demonstrate ZIMMA's superior performance in capturing distinct mediation mechanisms for both rare and abundant species compared to existing methods. Its application to human microbiome studies examining the effects of dietary intake and metabolic syndrome underscores its efficacy in identifying key microbial mediators, offering valuable insights and biological interpretation into the role of the microbiome in disease physiology and health sciences.

REFERENCES

AHMAD, M. A., KARAVETIAN, M., MOUBARECK, C. A., WAZZ, G., MAHDY, T. and VENEMA, K. (2023). Association of the gut microbiota with clinical variables in obese and lean Emirati subjects. *Front. Microbiol.* **14** 1182460.

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

BARON, R. M. and KENNY, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51** 1173.

BOONSTRA, P. S., BARBARO, R. P. and SEN, A. (2019). Default priors for the intercept parameter in logistic regressions. *Comput. Statist. Data Anal.* **133** 245–256. MR3926478 https://doi.org/10.1016/j.csda.2018.10.014

BÜRKNER, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* **80** 1–28.

ĆESIĆ, D., LUGOVIĆ MIHIĆ, L., OZRETIĆ, P., LOJKIĆ, I., BULJAN, M., ŠITUM, M., ZOVAK, M., VIDOVIĆ, D., MIJIĆ, A. et al. (2023). Association of gut lachnospiraceae and chronic spontaneous urticaria. *Life* **13** 1280.

CHEN, J., ZHANG, X., YANG, L. and ZHANG, L. (2023). GUniFrac: Generalized UniFrac Distances, Distance-Based Multivariate Methods and Feature-Based Univariate Methods for Microbiome Data Analysis R package version 1.8..

CHEN, L., REEVE, J., ZHANG, L., HUANG, S., WANG, X. and CHEN, J. (2018). GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* **6** e4600.

Key words and phrases. Microbiome, zero-inflation, causal mediation, Bayesian inference.

CHO, H., QU, Y., LIU, C., TANG, B., LYU, R., LIN, B. M., ROACH, J., AZCARATE-PERIL, M. A., AGUIAR RIBEIRO, A. et al. (2023). Comprehensive evaluation of methods for differential expression analysis of meta-transcriptomics data. *Brief. Bioinform.* **24** bbad279.

CLARKE, S. F., MURPHY, E. F., NILAWEERA, K., ROSS, P. R., SHANAHAN, F., O'TOOLE, P. W. and COTTER, P. D. (2012). The gut microbiota and its relationship to diet and obesity: New insights. *Gut Microbes* **3** 186–202.

DEFEUDIS, G., ROSSINI, M., KHAZRAI, Y., PIPICELLI, A., BRUCOLI, G., VENEZIANO, M., STROLLO, F., LAZIO, A.-S.-S.-L. S. G., BELLIA, A. et al. (2022). The gut microbiome as possible mediator of the beneficial effects of very low calorie ketogenic diet on type 2 diabetes and obesity: A narrative review. *Eat. Weight Disord. Anorex, Bulim. Obes.* **27** 2339–2346.

FEI, N. and ZHAO, L. (2013). An opportunistic pathogen isolated from the gut of an obese human causes obesity in germfree mice. *ISME J.* **7** 880–884.

FENG, C. X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *J. Stat. Distrib. Appl.* **8**.

FU, J., KOSLOVSKY, M. D., NEOPHYTOU, A. M. and VANNUCCI, M. (2023). A Bayesian joint model for compositional mediation effect selection in microbiome data. *Stat. Med.* **42** 2999–3015. MR4606382 https://doi.org/10.1002/sim.9764

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.

GOMES, A. C., HOFFMANN, C. and MOTA, J. F. (2020). Gut microbiota is associated with adiposity markers and probiotics may impact specific genera. *Eur. J. Nutr.* **59** 1751–1762.

HAMIDI, B., WALLACE, K. and ALEKSEYENKO, A. V. (2019). MODIMA, a method for multivariate omnibus distance mediation analysis, allows for integration of multivariate exposure–mediator–response relationships. *Genes* **10** 524.

HE, M., ZHAO, N. and SATTEN, G. A. (2024). MIDASim: A fast and simple simulator for realistic microbiome data. *Microbiome* **12** 135.

HEIMAN, M. L. and GREENWAY, F. L. (2016). A healthy gastrointestinal microbiome is dependent on dietary diversity. *Mol. Genet. Metab.* **5** 317–320.

HEINZE, G. and SCHEMPER, M. (2002). A solution to the problem of separation in logistic regression. *Stat. Med.* **21** 2409–2419.

HONG, Q., CHEN, G. and TANG, Z.-Z. (2023). PhyloMed: A phylogeny-based test of mediation effect in microbiome. *Genome Biol.* **24** 72.

HU, S., WANG, J., XU, Y., YANG, H., WANG, J., XUE, C., YAN, X. and SU, L. (2019). Anti-inflammation effects of fucosylated chondroitin sulphate from Acaudina molpadioides by altering gut microbiota in obese mice. *Food Funct.* **10** 1736–1746.

HUA, H., WAN, T., WENJUAN, W. and CRITS-CHRISTOPH, P. (2014). Structural zeroes and zero-inflated models. *Arch. Psychiatry* **26** 236.

IMAI, K., KEELE, L. and TINGLEY, D. (2010). A general approach to causal mediation analysis. *Psychol. Methods* **15** 309.

IMAI, K., KEELE, L., TINGLEY, D. and YAMAMOTO, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *Amer. Polit. Sci. Rev.* **105** 765–789.

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

ISHWARAN, H. and RAO, J. (2000). Bayesian nonparametric MCMC for large variable selection problems. Unpublished manuscript.

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

JARDON, K. M., CANFORA, E. E., GOOSSENS, G. H. and BLAAK, E. E. (2022). Dietary macronutrients and the gut microbiome: A precision nutrition approach to improve cardiometabolic health. *Gut* **71** 1214–1226.

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A* **186** 453–461. MR0017504 https://doi.org/10.1098/rspa.1946.0056

JIANG, M., LEE, S., O'MALLEY, A. J., STERN, Y. and LI, Z. (2023). A novel causal mediation analysis approach for zero-inflated mediators. *Stat. Med.* **42** 2061–2081. MR4594701 https://doi.org/10.1002/sim.9689

JINNO, C., WONG, B., KLÜNEMANN, M., HTOO, J., LI, X. and LIU, Y. (2023). Effects of supplementation of Bacillus amyloliquefaciens on performance, systemic immunity, and intestinal microbiota of weaned pigs experimentally infected with a pathogenic enterotoxigenic E. coli F18. *Front. Microbiol.* **14** 1101457.

JOSEPH, N., PAULSON, C., CORRADA BRAVO, H. and POP, M. (2013). Robust methods for differential abundance analysis in marker gene surveys. *Nat. Methods* **10** 1200–1202.

JUST, S., MONDOT, S., ECKER, J., WEGNER, K., RATH, E., GAU, L., STREIDL, T., HERY-ARNAUD, G., SCHMIDT, S. et al. (2018). The gut microbiota drives the impact of bile acids and fat source in diet on mouse metabolism. *Microbiome* **6** 134.

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.

KAUL, A., DAVIDOV, O. and PEDDADA, S. D. (2017). Structural zeros in high-dimensional data with applications to microbiome studies. *Biostatistics* **18** 422–433. MR3824758 https://doi.org/10.1093/biostatistics/kxw053

KOH, A., DE VADDER, F., KOVATCHEVA-DATCHARY, P. and BÄCKHED, F. (2016). From dietary fiber to host physiology: Short-chain fatty acids as key bacterial metabolites. *Cell* **165** 1332–1345.

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

LARSEN, J. M. (2017). The immune response to Prevotella bacteria in chronic inflammatory disease. *Immunology* **151** 363–374.

LIAO, C., TAYLOR, B. P., CECCARANI, C., FONTANA, E., AMORETTI, L. A., WRIGHT, R. J., GOMES, A. L., PELED, J. U., TAUR, Y. et al. (2021). Compilation of longitudinal microbiota data and hospitalome from hematopoietic cell transplantation patients. *Sci. Data* **8** 71.

LOVE, M. I., HUBER, W. and ANDERS, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15** 1–21.

MA, S., REN, B., MALLICK, H., MOON, Y. S., SCHWAGER, E., MAHARJAN, S., TICKLE, T. L., LU, Y., CARMODY, R. N. et al. (2021). A statistical model for describing and simulating microbial community profiles. *PLoS Comput. Biol.* **17** e1008913.

MACKINNON, D. (2012). *Introduction to Statistical Mediation Analysis*. Routledge, London.

MACKINNON, D. P., LOCKWOOD, C. M., HOFFMAN, J. M., WEST, S. G. and SHEETS, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods* **7** 83.

MANDAL, S., VAN TREUREN, W., WHITE, R. A., EGGESBØ, M., KNIGHT, R. and PEDDADA, S. D. (2015). Analysis of composition of microbiomes: A novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26** 27663.

MORJARIA, S., SCHLUTER, J., TAYLOR, B. P., LITTMANN, E. R., CARTER, R. A., FONTANA, E., PELED, J. U., VAN DEN BRINK, M. R., XAVIER, J. B. et al. (2019). Antibiotic-induced shifts in fecal microbiota density and composition during hematopoietic stem cell transplantation. *Infect. Immun.* **87** 10–1128.

MULLAHY, J. (1986). Specification and testing of some modified count data models. *J. Econometrics* **33** 341–365. MR0867980 https://doi.org/10.1016/0304-4076(86)90002-3

PAN, B., GUO, Q., CAI, J., CHEN, L., ZHAO, Z., SHEN, P. and WANG, Y. (2024). Investigating the causal impact of gut microbiota on arthritis via inflammatory proteins using Mendelian randomization. *Sci. Rep.* **14** 27433.

PEARL, J. (2014). Interpretation and identification of causal mediation. *Psychol. Methods* **19** 459.

PEARL, J. (2022). Direct and indirect effects. In *Probabilistic and Causal Inference*: *The Works of Judea Pearl* 373–392.

PELED, J. U., GOMES, A. L., DEVLIN, S. M., LITTMANN, E. R., TAUR, Y., SUNG, A. D., WEBER, D., HASHIMOTO, D., SLINGERLAND, A. E. et al. (2020). Microbiota as predictor of mortality in allogeneic hematopoietic-cell transplantation. *N. Engl. J. Med.* **382** 822–834.

PILLOW, J. and SCOTT, J. (2012). Fully Bayesian inference for neural models with negative-binomial spiking. In *Advances in Neural Information Processing Systems* (F. Pereira, C. J. Burges, L. Bottou and K. Q. Weinberger, eds.) **25**. Curran Associates, Red Hook.

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712 https://doi.org/10.1080/01621459.2013.829001

QIAO, X., LIU, R., TANG, X., PETERSON, C. B, JENQ, R. R and ZHANG, L. (2026). Supplement to "Decoding microbiome dual mediation: Introducing ZIMMA for enhanced zero-inflated data analysis." https://doi.org/10.1214/26-AOAS2170SUPPA, https://doi.org/10.1214/26-AOAS2170SUPPB

ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

SCHLUTER, J., PELED, J. U., TAYLOR, B. P., MARKEY, K. A., SMITH, M., TAUR, Y., NIEHUS, R., STAFFAS, A., DAI, A. et al. (2020). The gut microbiota is associated with immune cell dynamics in humans. *Nature* **588** 303–307.

SERRANO, D., POZZI, C., GUGLIETTA, S., FOSSO, B., SUPPA, M., GNAGNARELLA, P., CORSO, F., BELLERBA, F., MACIS, D. et al. (2021). Microbiome as mediator of diet on colorectal cancer risk: The role of vitamin D, markers of inflammation and adipokines. *Nutrients* **13** 363.

SHI, H., TER HORST, R., NIELEN, S., BLOEMENDAAL, M., JAEGER, M., JOOSTEN, I., KOENEN, H., JOOSTEN, L. A., SCHWEREN, L. J. et al. (2022). The gut microbiome as mediator between diet and its impact on immune function. *Sci. Rep.* **12** 5149.

SHIN, N.-R., WHON, T. W. and BAE, J.-W. (2015). Proteobacteria: Microbial signature of dysbiosis in gut microbiota. *Trends Biotechnol.* **33** 496–503.

SOHN, M. B. and LI, H. (2019). Compositional mediation analysis for microbiome studies. *Ann. Appl. Stat.* **13** 661–681. MR3937444 https://doi.org/10.1214/18-AOAS1210

TVEDEBRINK, T. (2010). Overdispersion in allelic counts and theta-correction in forensic genetics. *Theor. Popul. Biol.* **78** 200–210.

VACCA, M., CELANO, G., CALABRESE, F., PORTINCASA, P., GOBBETTI, M. and DE ANGELIS, M. (2020). The controversial role of human gut Lachnospiraceae. *Microorganisms* **8** 573.

VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*, 4th ed. Springer, New York. ISBN 0-387-95457-0.

WANG, C., AHN, J., TARPEY, T., YI, S. S., HAYES, R. B. and LI, H. (2023). A microbial causal mediation analytic tool for health disparity and applications in body mass index. *Microbiome* **11** 164.

WANG, C., HU, J., BLASER, M. J. and LI, H. (2020). Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics* **36** 347–355.

WU, G. D., CHEN, J., HOFFMANN, C., BITTINGER, K., CHEN, Y.-Y., KEILBAUGH, S. A., BEWTRA, M., KNIGHTS, D., WALTERS, W. A. et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334** 105–108.

WU, Q., O'MALLEY, J., DATTA, S., GHARAIBEH, R. Z., JOBIN, C., KARAGAS, M. R., COKER, M. O., HOEN, A. G., CHRISTENSEN, B. C. et al. (2022). MarZIC: A marginal mediation model for zero-inflated compositional mediators with applications to microbiome data. *Genes* **13**.

YANG, D. and XU, W. (2023). Estimation of mediation effect on zero-inflated microbiome mediators. *Mathematics* **11**.

YANG, L. and CHEN, J. (2022). A comprehensive evaluation of microbial differential abundance analysis methods: Current status and potential solutions. *Microbiome* **10** 130.

YATES, F. (1934). Contingency tables involving small numbers and the χ 2 test. *Suppl. J. R. Stat. Soc.* **1** 217–235.

YUE, Y. and HU, Y.-J. (2022). A new approach to testing mediation of the microbiome at both the community and individual taxon levels. *Bioinformatics* **38** 3173–3180.

ZHANG, H., CHEN, J., FENG, Y., WANG, C., LI, H. and LIU, L. (2021). Mediation effect selection in high-dimensional and compositional microbiome data. *Stat. Med.* **40** 885–896. MR4201109 https://doi.org/10.1002/sim.8808

ZHANG, J., WEI, Z. and CHEN, J. (2018). A distance-based approach for testing the mediation effect of the human microbiome. *Bioinformatics* **34** 1875–1883.

ZHAO, L., ZHANG, F., DING, X., WU, G., LAM, Y. Y., WANG, X., FU, H., XUE, X., LU, C. et al. (2018). Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. *Science* **359** 1151–1156.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 986–1009
https://doi.org/10.1214/26-AOAS2160

A PRINCIPAL SUBMANIFOLD-BASED APPROACH FOR CLUSTERING AND MULTISCALE RNA CORRECTION

BY MENGHAO WU[a] AND ZHIGANG YAO[b]

Department of Statistics and Data Science, National University of Singapore, [a]*menghao.wu@u.nus.edu,* [b]*zhigang.yao@nus.edu.sg*

RNA structure determination is essential for understanding its biological functions. However, the reconstruction process often faces challenges, such as atomic clashes, which can lead to inaccurate models. To address these challenges, we introduce the principal submanifold (PSM) approach for analyzing RNA data on a torus. This method provides an accurate, low-dimensional feature representation, overcoming the limitations of previous torus-based methods. By combining PSM with DBSCAN, we propose a novel clustering technique, the principal submanifold-based DBSCAN (PSM-DBSCAN). Our approach achieves superior clustering accuracy and increased robustness to noise. Additionally, we apply this new method for multiscale corrections, effectively resolving RNA backbone clashes at both microscopic and mesoscopic scales. Extensive simulations and comparative studies highlight the enhanced precision and scalability of our method, demonstrating significant improvements over existing approaches. The proposed methodology offers a robust foundation for correcting complex RNA structures and has broad implications for applications in structural biology and bioinformatics.

REFERENCES

BANK, P. D. (1971). Protein data bank. *Nat*., *New Biol*. **233** 10–1038.

BATSANOV, S. S. (2001). Van der Waals radii of elements. *Inorg*. *Mater.* **37** 871–885.

CHEN, L.-L. (2020). The expanding regulatory mechanisms and cellular functions of circular RNAs. *Nat*. *Rev*., *Mol*. *Cell Biol*. **21** 475–490.

CHOU, F.-C., SRIPAKDEEVONG, P., DIBROV, S. M., HERMANN, T. and DAS, R. (2013). Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nat*. *Methods* **10** 74–76.

DONOHO, D. L. and GRIMES, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **100** 5591–5596. MR1981019 https://doi.org/10.1073/pnas.1031596100

DOUDNA, J. A. and CECH, T. R. (2002). The chemical repertoire of natural ribozymes. *Nature* **418** 222–228.

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

ELTZNER, B., HUCKEMANN, S. and MARDIA, K. V. (2018). Torus principal component analysis with applications to RNA structure. *Ann*. *Appl*. *Stat*. **12** 1332–1359. MR3834306 https://doi.org/10.1214/17-AOAS1115

ESTER, M., KRIEGEL, H.-P., SANDER, J., XU, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* **96** 226–231.

FEFFERMAN, C., MITTER, S. and NARAYANAN, H. (2016). Testing the manifold hypothesis. *J*. *Amer*. *Math*. *Soc*. **29** 983–1049. MR3522608 https://doi.org/10.1090/jams/852

JAIN, S., RICHARDSON, D. C. and RICHARDSON, J. S. (2015). Computational methods for RNA structure validation and improvement. In *Methods in Enzymology* **558** 181–212. Elsevier, Amsterdam.

JINEK, M., JIANG, F., TAYLOR, D. W., STERNBERG, S. H., KAYA, E., MA, E., ANDERS, C., HAUER, M., ZHOU, K. et al. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343** 1247997.

JUNG, S., DRYDEN, I. L. and MARRON, J. S. (2012). Analysis of principal nested spheres. *Biometrika* **99** 551–568. MR2966769 https://doi.org/10.1093/biomet/ass022

KIEFT, J. S., ZHOU, K., GRECH, A., JUBIN, R. and DOUDNA, J. A. (2002). Crystal structure of an RNA tertiary domain essential to HCV IRES-mediated translation initiation. *Nat*. *Struct*. *Biol*. **9** 370–374.

Key words and phrases. Geometric statistics, principal submanifold, dimensionality reduction, high-dimensional clustering, clash correction.

LILLEY, D. M. (2000). Structures of helical junctions in nucleic acids. *Q. Rev. Biophys.* **33** 109–159.
LU, S., TANG, Y., YIN, S. and SUN, L. (2024). RNA structure: Implications in viral infections and neurodegenerative diseases. *Adv. Biotechnol.* **2** 3.
MARDIA, K. V. (2013). Statistical approaches to three key challenges in protein structural bioinformatics. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **62** 487–514. MR3060628 https://doi.org/10.1111/rssc.12003
MARDIA, K. V., WIECHERS, H., ELTZNER, B. and HUCKEMANN, S. F. (2022). Principal component analysis and clustering on manifolds. *J. Multivariate Anal.* **188** Paper No. 104862, 21. MR4353867 https://doi.org/10.1016/j.jmva.2021.104862
MCINNES, L., HEALY, J. and MELVILLE, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint. Available at arXiv:1802.03426.
MOORE, P. B. (1998). The three-dimensional structure of the ribosome and its components. *Annu. Rev. Biophys. Biomol. Struct.* **27** 35–58.
MURRAY, L. J., ARENDALL, W. B. III, RICHARDSON, D. C. and RICHARDSON, J. S. (2003). RNA backbone is rotameric. *Proc. Natl. Acad. Sci. USA* **100** 13904–13909.
PANARETOS, V. M., PHAM, T. and YAO, Z. (2014). Principal flows. *J. Amer. Statist. Assoc.* **109** 424–436. MR3180574 https://doi.org/10.1080/01621459.2013.849199
ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326.
SERGANOV, A. and NUDLER, E. (2013). A decade of riboswitches. *Cell* **152** 17–24.
SHEN, L. X., CAI, Z. and TINOCO, I. JR (1995). RNA structure at high resolution. *FASEB J.* **9** 1023–1033.
SHIVASHANKAR, G. (2002). Mesoscopic biology. *Pramāna* **58** 439–442.
SPONER, J., BUSSI, G., KREPL, M., BANÁŠ, P., BOTTARO, S., CUNHA, R. A., GIL-LEY, A., PINAMONTI, G., POBLETE, S. et al. (2018). RNA structural dynamics as captured by molecular simulations: A comprehensive overview. *Chem. Rev.* **118** 4177–4338.
TENENBAUM, J. B., SILVA, V. D. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.
VAN DER MAATEN, L. and HINTON, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** 2579–2605.
WIECHERS, H., ELTZNER, B., MARDIA, K. V. and HUCKEMANN, S. F. (2023). Learning torus PCA-based classification for multiscale RNA correction with application to SARS-CoV-2. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **72** 271–293. MR4719276 https://doi.org/10.1093/jrsssc/qlad004
WU, M. and YAO, Z. (2026). Supplement to "A principal submanifold-based approach for clustering and multiscale RNA correction." https://doi.org/10.1214/26-AOAS2160SUPPA, https://doi.org/10.1214/26-AOAS2160SUPPB
YAO, Z., ELTZNER, B. and PHAM, T. (2026). Principal sub-manifolds. *Statist. Sinica* **36** 1–41.
YAO, Z. and XIA, Y. (2025). Manifold fitting under unbounded noise. *J. Mach. Learn. Res.* **26** Paper No. [45], 55. MR4896094
YAO, Z., XIA, Y., TRAN, D. V. and ZHANG, Z. (2023). Hunting principal sub-manifolds: New theories and methods. Technical Report.
ZHU, Y., ZHU, L., WANG, X. and JIN, H. (2022). RNA-based therapeutics: An overview and prospectus. *Cell Death Dis.* **13** 644.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1010–1032
https://doi.org/10.1214/26-AOAS2165

DOMAIN-AWARE MATRIX COMPLETION FOR PHENOTYPE IMPUTATION USING ELECTRONIC HEALTH RECORD DATA WITH APPLICATIONS IN GENOMIC RESEARCH

BY HANQING WU[1,a], CUE HYUNKYU LEE[2,b], NAJMEH ABIRI[3,d] AND IULIANA IONITA-LAZA[1,2,c]

[1]*Department of Statistics, Lund University,* [a]*hanqing.wu@stat.lu.se*
[2]*Department of Biostatistics, Columbia University,* [b]*hl3565@cumc.columbia.edu,* [c]*ii2135@cumc.columbia.edu*
[3]*Department of Computer Science and Media Technology, Malmö University,* [d]*najmeh.abiri@mau.se*

Large-scale biobanks and electronic health records (EHR) offer great opportunities for next-generation genetic studies. However, missing phenotype data is a pervasive feature of EHR, leading to low power of such studies. One promising solution is prediction-powered inference, where statistical or machine learning models are employed to impute phenotypes prior to performing genetic analyses. Although many such methods exist, they tend to be generic and do not incorporate domain-aware knowledge to optimize their performance for downstream genetic analyses. We propose a novel matrix completion method, covImpute, which, unlike generic matrix completion methods such as softImpute, incorporates external information in the form of a genetic covariance matrix among phenotypic features and imputes missing entries with latent genetic components. We compare covImpute with existing methods, including a domain-aware liability threshold model LTPI, and generic softImpute and deep learning autoencoder models in simulations under different missingness mechanisms with respect to power in downstream genetic analyses. In applications to several diseases in UK Biobank, we show that genetically informed methods, such as covImpute and LTPI, can perform substantially better in terms of power of genetic association studies relative to generic imputation models currently in use. Moreover, compared to LTPI, covImpute's flexible framework for incorporating external covariance information provides a more general approach with applicability beyond genetics.

REFERENCES

ALAYA, M. Z. and KLOPP, O. (2019). Collective matrix completion. *J. Mach. Learn. Res.* **20** Paper No. 148. MR4030162

ALLEN, G. I. and TIBSHIRANI, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *Ann. Appl. Stat.* **4** 764–790. MR2758420 https://doi.org/10.1214/09-AOAS314

AN, U. L., PAZOKITOROUDI, A., ALVAREZ, M., HUANG, L. Y., BACANU, S., SCHORK, A. J., KENDLER, K., PAJUKANTA, P., FLINT, J. et al. (2023). Deep learning-based phenotype imputation on population-scale biobank data increases genetic discoveries. *Nat. Genet.* **55** 2269–2272.

BAGHERIAN, M., KIM, R. B., JIANG, C., SARTOR, M. A., DERKSEN, H. and NAJARIAN, K. (2021). Coupled matrix–matrix and coupled tensor–matrix completion methods for predicting drug–target interactions. *Brief. Bioinform.* **22** 2161–2171.

BASTARACHE, L. (2021). Using phecodes for research with the electronic health record: From PheWAS to PheRS. *Annu. Rev. Biomed. Data Sci.* **4** 1–19.

BEAULIEU-JONES, B. K., LAVAGE, D. R., SNYDER, J. W., MOORE, J. H., PENDERGRASS, S. A. and BAUER, C. R. (2018). Characterizing and managing missing structured data in electronic health records: Data analysis. *JMIR Med. Inform.* **6** e8960.

BULIK-SULLIVAN, B., FINUCANE, H. K., ANTTILA, V., GUSEV, A., DAY, F. R., LOH, P. R., DUNCAN, L., PERRY, J. R. B., PATTERSON, N. et al. (2015a). An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47** 1236.

Key words and phrases. Electronic health records (EHR), matrix completion method, genetic liability matrix, phenotype imputation, biobanks, GWAS.

BULIK-SULLIVAN, B. K., LOH, P. R., FINUCANE, H. K., RIPKE, S., YANG, J., PATTERSON, N., DALY, M. J., PRICE, A. L., NEALE, B. M. et al. (2015b). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47** 291.

CAI, D., HE, X., HAN, J. and HUANG, T. S. (2010). Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 1548–1560.

CAI, N. (2023). MDDImpute. GitHub.

CHEN, Y., CHENG, L. and WU, Y.-C. (2023). Bayesian low-rank matrix completion with dual-graph embedding: Prior analysis and tuning-free inference. *Signal Process.* **204** 108826.

CHIANG, K.-Y., HSIEH, C.-J. and DHILLON, I. S. (2015). Matrix completion with noisy side information. *Adv. Neural Inf. Process. Syst.* **28**.

DAHL, A., IOTCHKOVA, V., BAUD, A., JOHANSSON, A., GYLLENSTEN, U., SORANZO, N., MOTT, R., KRANIS, A. and MARCHINI, J. (2016). A multiple-phenotype imputation method for genetic studies. *Nat. Genet.* **48** 466.

DAHL, A., THOMPSON, M., AN, U. L., KREBS, M., APPADURAI, V., BORDER, R., BACANU, S. A., WERGE, T., FLINT, J. et al. (2023). Phenotype integration improves power and preserves specificity in biobank-based genetic studies of major depressive disorder. *Nat. Genet.* **55** 2082–2093.

DAVENPORT, M. A., PLAN, Y., VAN DEN BERG, E. and WOOTTERS, M. (2014). 1-bit matrix completion. *Inf. Inference* **3** 189–223. MR3311452 https://doi.org/10.1093/imaiai/iau006

FAN, J., LI, X. C., CROVELLA, M. and LEISERSON, M. D. M. (2020). Matrix (factorization) reloaded: Flexible methods for imputing genetic interactions with cross-species and side information. *Bioinformatics* **36** i866–i874.

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, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA J. Numer. Anal.* **22** 329–343. MR1918653 https://doi.org/10.1093/imanum/22.3.329

HUJOEL, M. L. A., GAZAL, S., LOH, P. R., PATTERSON, N. and PRICE, A. L. (2020). Liability threshold modeling of case-control status and family history of disease increases association power. *Nat. Genet.* **52** 541.

JAIN, P. and DHILLON, I. S. (2013). Provable Inductive Matrix Completion. Available at arXiv:1306.0626.

JAIN, S., CHOUZENOUX, E., KUMAR, K. and MAJUMDAR, A. (2023). Graph regularized probabilistic matrix factorization for drug-drug interactions prediction. *IEEE J. Biomed. Health Inform.* **27** 2565–2574.

JANSEN, I. E., SAVAGE, J. E., WATANABE, K., BRYOIS, J., WILLIAMS, D. M., STEINBERG, S., SEALOCK, J., KARLSSON, I. K., HÄGG, S., ATHANASIU, L., VOYLE, N., PROITSI, P., WITOELAR, A., STRINGER, S., AARSLAND, D., ALMDAHL, I. S., ANDERSEN, F., BERGH, S., BETTELLA, F., BJORNSSON, S., BRÆKHUS, A., BRÅTHEN, G., DE LEEUW, C., DESIKAN, R. S., DJUROVIC, S., DUMITRESCU, L., FLADBY, T., HOHMAN, T. J., JONSSON, P. V., KIDDLE, S. J., RONGVE, A., SALTVEDT, I., SANDO, S. B., SELBÆK, G., SHOAI, M., SKENE, N. G., SNAEDAL, J., STORDAL, E., ULSTEIN, I. D., WANG, Y. P., WHITE, L. R., HARDY, J., HJERLING-LEFFLER, J., SULLIVAN, P. F., VAN DER FLIER, W. M., DOBSON, R., DAVIS, L. K., STEFANSSON, H., STEFANSSON, K., PEDERSEN, N. L., RIPKE, S., ANDREASSEN, O. A. and POSTHUMA, D. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51** 404.

KARCZEWSKI, K. J., GUPTA, R., KANAI, M., LU, W., TSUO, K., WANG, Y., WALTERS, R. K., TURLEY, P., CALLIER, S. et al. (2024). Pan-UK Biobank GWAS improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects. *MedRxiv*.

LEE, C. (2024). LTPI software. https://github.com/cuelee/LTPI. Accessed: 2024-10-22.

LEE, C. H., KHAN, A., WANG, C., WENG, C., BUXBAUM, J. D., KIRYLUK, K. and IONITA-LAZA, I. (2025). Liability threshold model-based disease risk prediction based on electronic health record phenotypes. *Nat. Genet.* **57** 2872–2881.

MAHALANOBIS, P. C. (2018). Reprint of: On the generalised distance in statistics. *Sankhya A* **80** 1–7.

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

MBATCHOU, J., BARNARD, L., BACKMAN, J., MARCKETTA, A., KOSMICKI, J. A., ZIYATDINOV, A., BENNER, C., O'DUSHLAINE, C., BARBER, M. et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53** 1097.

MBATCHOU, M. J. (2021). Recommendations for UK Biobank analysis. Available at https://rgcgithub.github.io/regenie/recommendations/.

MCCAW, Z. R., GAO, J. H., LIN, X. H. and GRONSBELL, J. (2024). Synthetic surrogates improve power for genome-wide association studies of partially missing phenotypes in population biobanks. *Nat. Genet.* **56**.

MCCAW, Z. R., LANE, J. M., SAXENA, R., REDLINE, S. and LIN, X. (2020). Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics* **76** 1262–1272. MR4186840 https://doi.org/10.1111/biom.13214

MIAO, J. C., WU, Y. X., SUN, Z. X., MIAO, X. R., LU, T. Y., ZHAO, J. W. and LU, Q. S. (2024). Valid inference for machine learning-assisted genome-wide association studies. *Nat. Genet.* **56** 2361–2369.

NATARAJAN, N. and DHILLON, I. S. (2014). Inductive matrix completion for predicting gene–disease associations. *Bioinformatics* **30** i60–i68.

NEALE, B. M. (2014). Liability threshold models. In *Statistical Human Genetics*: *Methods and Protocols* Wiley, Inc, Hoboken, NJ.

NESTEROV, Y. (2018). *Lectures on Convex Optimization*. *Springer Optimization and Its Applications* **137**. Springer, Cham. MR3839649 https://doi.org/10.1007/978-3-319-91578-4

RAO, N., YU, H.-F., RAVIKUMAR, P. K. and DHILLON, I. S. (2015). Collaborative filtering with graph information: Consistency and scalable methods. *Adv. Neural Inf. Process. Syst.* **28**.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581

SHAN, H. and BANERJEE, A. (2010). Generalized probabilistic matrix factorizations for collaborative filtering. In 2010 *IEEE International Conference on Data Mining* 1025–1030. IEEE Press, New York.

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

SUDLOW, C., GALLACHER, J., ALLEN, N., BERAL, V., BURTON, P., DANESH, J., DOWNEY, P., ELLIOTT, P., GREEN, J. et al. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**.

TALIUN, S. A. G., VANDEHAAR, P., BOUGHTON, A. P., WELCH, R. P., TALIUN, D., SCHMIDT, E. M., ZHOU, W., NIELSEN, J. B., WILLER, C. J. et al. (2020). Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat. Genet.* **52** 550–552.

TAN, J., ZHANG, Y., HONG, C., CAI, T. T., CAI, T. and ZHANG, A. R. (2025). Integrated analysis for electronic health records with structured and sporadic missingness. Available at arXiv:2506.09208.

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** 1–67.

VERMA, A., HUFFMAN, J. E., RODRIGUEZ, A., CONERY, M., LIU, M. L., HO, Y. L., KIM, Y., HEISE, D. A., GUARE, L. et al. (2024). Diversity and scale: Genetic architecture of 2068 traits in the VA million veteran program. *Science* **385**.

WANG, K., KIM, N., BAGHERIAN, M., LI, K., CHOU, E., COLACINO, J. A., DOLINOY, D. C. and SARTOR, M. A. (2024). Gene target prediction of environmental chemicals using coupled matrix–matrix completion. *Environ. Sci. Technol.* **58** 5889–5898.

WANG, Y., YANG, Y., WANG, K., GAO, S. and LIAO, X. (2025). Matrix Completion with Graph Information: a Provable Nonconvex Optimization Approach. Available at arXiv:2502.08536.

WEI, W. Q. and DENNY, J. C. (2015). Extracting research-quality phenotypes from electronic health records to support precision medicine. *Gen. Med.* **7**.

WU, H., LEE, C. H., ABIRI, N. and IONITA-LAZA, I. (2026). Supplement to "Domain-aware matrix completion for phenotype imputation using Electronic Health Record data with applications in genomic research." https://doi.org/10.1214/26-AOAS2165SUPPA, https://doi.org/10.1214/26-AOAS2165SUPPB

YOON, J., JORDON, J. and VAN DER SCHAAR, M. (2018). GAIN: Missing data imputation using generative adversarial nets. *Int. Conf. Appl. Mach. Learn.* **80**.

ZHANG, Z. and ZHAO, K. (2013). Low-rank matrix approximation with manifold regularization. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 1717–1729.

ZHOU, T., SHAN, H., BANERJEE, A. and SAPIRO, G. (2012). Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In *Proceedings of the* 2012 *SIAM International Conference on Data Mining* 403–414. Society for Industrial and Applied Mathematics.

ZHOU, W., KANAI, M., WU, K. H. H., RASHEED, H., TSUO, K., HIRBO, J. B., WANG, Y., BHATTACHARYA, A., ZHAO, H. L. et al. (2022). Global Biobank meta-analysis initiative: Powering genetic discovery across human disease. *Cell Genom.* **2** 100192.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1033–1057
https://doi.org/10.1214/26-AOAS2166

TARGETED MAXIMUM LIKELIHOOD ESTIMATION FOR INTEGRAL PROJECTION MODELS IN POPULATION ECOLOGY

BY YUNZHE ZHOU[1,a] AND GILES HOOKER[2,b]

[1]*Department of Biostatistics, University of California, Berkeley,* [a]*ztzyz615@berkeley.edu*
[2]*Department of Statistics and Data Science, University of Pennsylvania,* [b]*ghooker@wharton.upenn.edu*

In population ecology, integral projection models (IPMs) are widely used to study population growth and the dynamics of population structure (e.g., age and size distributions). These models typically use data on the growth, survival, and reproduction of marked individuals to parameterize models for each demographic rate. The resulting models can be used to predict changes in the population from one time point to the next and to predict long-term properties such as long-term population growth rate, the sensitivity of that growth rate to environmental factors and to changes in demographic rates, and the variation in lifetime outcomes among individuals. These quantities must be inferred from the model, and we often lack any way to ground-truth their plausibility directly from the available data. They nonetheless reflect key ecological processes with significant management and policy implications. Building IPMs requires us to develop sub-models for individual fates over the next time step—Did they survive? How much did they grow or shrink? Did they reproduce?—conditional on their initial state as well as on environmental covariates. This must be done in a manner that gives the most precise possible estimates, and most accurate uncertainty quantification, for the long-term model properties that we are interested in predicting. These models include three core demographic submodels—for growth, survival, and reproduction rates—to describe how individuals change from one time point to the next.

Targeted maximum likelihood estimation (TMLE) methods are particularly well suited to a situation in which we are largely interested in estimation and inference on quantities derived from models. These methods build machine learning-based models that estimate the probability distribution from which the empirical data were drawn, and then the user specifies a *target of inference* as a function of this distribution. An initial estimate for the distribution is then modified by tilting in the direction of the influence function to both de-bias the parameter estimate and provide more accurate inference. In this paper we employ TMLE to develop robust and efficient estimators for quantities derived from a fitted IPM as targets of interest, with a particular focus on long-term stable population growth, its elasticity to fecundity, and the expected growth rate under year-specific covariates. Mathematically, we derive the influence functions for these targets of influence and formulate and use these to update our model so as to remove bias from our estimates. Empirically, we conduct extensive simulations and demonstrate our method's efficacy using empirical data from a long-term study of plant communities on the Idaho steppe and from experimental rotifer populations.

REFERENCES

ADLER, P. B., DALGLEISH, H. J. and ELLNER, S. P. (2012). Forecasting plant community impacts of climate variability and change: When do competitive interactions matter? *J. Ecol.* **100** 478–487.

ADLER, P. B., ELLNER, S. P. and LEVINE, J. M. (2010). Coexistence of perennial plants: An embarrassment of niches. *Ecol. Lett.* **13** 1019–1029.

Key words and phrases. Integral projection model, targeted maximum likelihood estimation, influence function.

ADLER, P. B., KLEINHESSELINK, A., HOOKER, G., TAYLOR, J. B., TELLER, B. and ELLNER, S. P. (2018). Weak interspecific interactions in a sagebrush steppe? Conflicting evidence from observations and experiments. *Ecology* **99** 1621–1632.

ADLER, P. B., KLEINHESSELINK, A., HOOKER, G., TAYLOR, J. B., TELLER, B. and ELLNER, S. P. (2019). Data from: Weak interspecific interactions in a sagebrush steppe? Conflicting evidence from observations and experiments. Dataset. https://doi.org/10.5061/dryad.96dn293

BENKESER, D. and VAN DER LAAN, M. (2016). The highly adaptive lasso estimator. In 2016 *IEEE International Conference on Data Science and Advanced Analytics* (*DSAA*) 689–696. IEEE Press, New York.

BOCK, M. J., JARVIS, G. C., COREY, E. L., STONE, E. E. and GRIBBLE, K. E. (2019). Maternal age alters offspring lifespan, fitness, and lifespan extension under caloric restriction. *Sci. Rep.* **9** 3138.

BRIGGS, J., DABBS, K., HOLM, M., LUBBEN, J., REBARBER, R., TENHUMBERG, B. and RISER-ESPINOZA, D. (2010). Structured population dynamics: An introduction to integral modeling. *Math. Mag.* **83** 243–257. MR2732318 https://doi.org/10.4169/002557010X521778

CASWELL, H. (2001). *Matrix Population Models*: *Construction*, *Analysis*, *and Interpretation*. Sinauer, Sunderland, Mass.

CHILDS, D. Z., REES, M., ROSE, K. E., GRUBB, P. J. and ELLNER, S. P. (2004). Evolution of size-dependent flowering in a variable environment: Construction and analysis of a stochastic integral projection model. *Proc. R. Soc. Lond., B Biol. Sci.* **271** 425–434.

COMPAGNONI, A., BIBIAN, A. J., OCHOCKI, B. M., ROGERS, H. S., SCHULTZ, E. L., SNECK, M. E., ELDERD, B. D., ILER, A. M., INOUYE, D. W. et al. (2016). The effect of demographic correlations on the stochastic population dynamics of perennial plants. *Ecol. Monogr.* **86** 480–494.

CRONE, E. E. (2016). Contrasting effects of spatial heterogeneity and environmental stochasticity on population dynamics of a perennial wildflower. *J. Ecol.* **104** 281–291.

DOAK, D. F., WADDLE, E., LANGENDORF, R. E., LOUTHAN, A. M., ISABELLE CHARDON, N., DIBNER, R. R., KEINATH, D. A., LOMBARDI, E., STEENBOCK, C. et al. (2021). A critical comparison of integral projection and matrix projection models for demographic analysis. *Ecol. Monogr.* **91** e01447. https://doi.org/10.1002/ecm.1447

DREES, T., OCHOCKI, B. M., COLLINS, S. L. and MILLER, T. E. (2023). Demography and dispersal at a grass-shrub ecotone: A spatial integral projection model for woody plant encroachment. *Ecol. Monogr.* e1574.

EASTERLING, M. R., ELLNER, S. P. and DIXON, P. M. (2000). Size-specific sensitivity: Applying a new structured population model. *Ecology* **81** 694–708.

ELLNER, S. P., CHILDS, D. Z. and REES, M. (2016). *Data-Driven Modelling of Structured Populations*: *A Practical Guide to the Integral Projection Model*. *Lecture Notes on Mathematical Modelling in the Life Sciences*. Springer, New York. MR3496931 https://doi.org/10.1007/978-3-319-28893-2

ELLNER, S. P. and REES, M. (2006). Integral projection models for species with complex demography. *Amer. Nat.* **167** 410–428.

FARRELL, M. H., LIANG, T. and MISRA, S. (2021). Deep neural networks for estimation and inference. *Econometrica* **89** 181–213. MR4220387 https://doi.org/10.3982/ecta16901

GELFAND, A. E., GHOSH, S. and CLARK, J. S. (2013). Scaling integral projection models for analyzing size demography. *Statist. Sci.* **28** 641–658. MR3161591 https://doi.org/10.1214/13-STS444

GROSS, K., MORRIS, W. F., WOLOSIN, M. S. and DOAK, D. F. (2006). Modeling vital rates improves estimation of population projection matrices. *Popul. Ecol.* **48** 79–89.

HERNÁNDEZ, C. M., VAN DAALEN, S. F., CASWELL, H., NEUBERT, M. G. and GRIBBLE, K. E. (2020). A demographic and evolutionary analysis of maternal effect senescence. *Proc. Natl. Acad. Sci. USA* **117** 16431–16437.

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

LAUENROTH, W. K. and ADLER, P. B. (2008). Demography of perennial grassland plants: Survival, life expectancy and life span. *J. Ecol.* **96** 1023–1032.

LOUTHAN, A. M., KEIGHRON, M., KIEKEBUSCH, E., CAYTON, H., TERANDO, A. and MORRIS, W. F. (2022). Climate change weakens the impact of disturbance interval on the growth rate of natural populations of Venus flytrap. *Ecol. Monogr.* **92** e1528.

MEROW, C., DAHLGREN, J. P., METCALF, C. J. E., CHILDS, D. Z., EVANS, M. E., JONGEJANS, E., RECORD, S., REES, M., SALGUERO-GÓMEZ, R. et al. (2014). Advancing population ecology with integral projection models: A practical guide. *Methods Ecol. Evol.* **5** 99–110.

METCALF, C. J. E., GRAHAM, A. L., MARTINEZ-BAKKER, M. and CHILDS, D. Z. (2016). Opportunities and challenges of integral projection models for modelling host–parasite dynamics. *J. Anim. Ecol.* **85** 343–355.

METCALF, J. C., ROSE, K. E. and REES, M. (2003). Evolutionary demography of monocarpic perennials. *Trends Ecol. Evol.* **18** 471–480.

MILLER, T. E. X. and ELLNER, S. P. (2025). My, how you've grown: A practical guide to modeling size transitions for Integral Projection Model (IPM) applications. *Ecology* **106** e70088.

MORRIS, W. F. and DOAK, D. F. (2002). *Quantitative Conservation Biology*: *Theory and Practice of Population Viability Analysis*. Sinauer, Sunderland, Mass.

OZGUL, A., CHILDS, D. Z., OLI, M. K., ARMITAGE, K. B., BLUMSTEIN, D. T., OLSON, L. E., TULJAPURKAR, S. and COULSON, T. (2010). Coupled dynamics of body mass and population growth in response to environmental change. *Nature* **466** 482–485.

PETERSEN, K. B., PEDERSEN, M. S. et al. (2008). The matrix cookbook. Technical Univ. Denmark 7 510.

REES, M. and ELLNER, S. P. (2009). Integral projection models for populations in temporally varying environments. *Ecol. Monogr.* **79** 575–594.

REES, M. and ELLNER, S. P. (2016). Evolving integral projection models: Evolutionary demography meets eco-evolutionary dynamics. *Methods Ecol. Evol.* **7** 157–170.

ROSENBLUM, M. and VAN DER LAAN, M. J. (2010). Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *Int. J. Biostat.* **6** Art. 13, 40. MR2609787 https://doi.org/10.2202/1557-4679.1138

SHORACK, G. R. and WELLNER, J. A. (2009). *Empirical Processes with Applications to Statistics*. *Classics in Applied Mathematics* **59**. SIAM, Philadelphia, PA. Reprint of the 1986 original [MR0838963]. MR3396731 https://doi.org/10.1137/1.9780898719017.ch1

SHRIVER, R. K., CUTLER, K. and DOAK, D. F. (2012). Comparative demography of an epiphytic lichen: Support for general life history patterns and solutions to common problems in demographic parameter estimation. *Oecologia* **170** 137–146.

SNYDER, R. E. and ELLNER, S. P. (2018). Pluck or luck: Does trait variation or chance drive variation in lifetime reproductive success? *Amer. Nat.* **191** E90–E107.

SNYDER, R. E., ELLNER, S. P. and HOOKER, G. (2021). Time and chance: Using age partitioning to understand how luck drives variation in reproductive success. *Amer. Nat.* **197** E110–E128.

SPREAFICO, M. (2024). Positivity violations in marginal structural survival models with time-dependent confounding: a simulation study on IPTW-estimator performance. arXiv preprint. Available at arXiv:2403.19606.

TREDENNICK, A. T., HOOKER, G., ELLNER, S. P. and ADLER, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology* **102**. e03336.

TREDENNICK, A. T., TELLER, B. J., ADLER, P. B., HOOKER, G. and ELLNER, S. P. (2018). Size-by-environment interactions: A neglected dimension of species' responses to environmental variation. *Ecol. Lett.* **21** 1757–1770.

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 ROSE, S. (2018). *Targeted Learning in Data Science*: *Causal Inference for Complex Longitudinal Studies*. *Springer Series in Statistics*. Springer, Cham. MR3791826 https://doi.org/10.1007/978-3-319-65304-4

VAN DER LAAN, M. J., ROSE, S. and ROSENBLUM, M. (2011). Robust analysis of RCTs using generalized linear models. In *Targeted Learning*: *Causal Inference for Observational and Experimental Data*. *Springer Ser. Statist.* 187–199. Springer, New York. MR2867123 https://doi.org/10.1007/978-1-4419-9782-1_11

VAN DER LAAN, M. J., ROSE, S., ZHENG, W. and VAN DER LAAN, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*: *Causal Inference for Observational and Experimental Data*. *Springer Ser. Statist.* 459–474. Springer, New York. MR2867139 https://doi.org/10.1007/978-1-4419-9782-1_27

WECKERLY, F. W. (2001). Matrix population models: Construction analysis and interpretation.

WHITE, J. W., NICKOLS, K. J., MALONE, D., CARR, M. H., STARR, R. M., CORDOLEANI, F., BASKETT, M. L., HASTINGS, A. and BOTSFORD, L. W. (2016). Fitting state-space integral projection models to size-structured time series data to estimate unknown parameters. *Ecol. Appl.* **26** 2677–2694.

WILBER, M. Q., LANGWIG, K. E., KILPATRICK, A. M., MCCALLUM, H. I. and BRIGGS, C. J. (2016). Integral projection models for host–parasite systems with an application to amphibian chytrid fungus. *Methods Ecol. Evol.* **7** 1182–1194.

ZHOU, Y. and HOOKER, G. (2026). Supplement to "Targeted maximum likelihood estimation for integral projection models in population ecology." https://doi.org/10.1214/26-AOAS2166SUPP

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1058–1077
https://doi.org/10.1214/26-AOAS2138

BAYESIAN SELECTION AND REFITTING OF REFERENCE MUTATIONAL SIGNATURES IN CANCER GENOMICS

BY MIN HUA[1,2,a] AND BIN ZHU[2,b]

[1]*Department of Mathematics and Statistics, Beck College of Sciences and Mathematics, Arkansas State University,* [a]*mhua@astate.edu*

[2]*Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health,* [b]*bin.zhu@nih.gov*

Somatic mutations, which accumulate in cells after conception, drive cancer development. Their characteristic patterns, namely mutational signatures, reflect the underlying mutational processes and have provided valuable insights into cancer etiology, evolution and therapeutic strategies. While nonnegative matrix factorization (NMF) is commonly used to infer de novo mutational signatures and their activities, it requires large datasets for reliable estimation. When the sample size is limited, signature refitting is typically used, which estimates the signature activities using a set of reference signatures derived from external studies. However, current signature refitting methods often use the full list of reference signatures, leading to overfitting and compromised interpretability and accuracy. Despite its importance, the problem of selecting an appropriate subset of reference signatures received little attention. We proposed BayesSigRefitting, a Bayesian model selection framework to select an optimal subset of reference signatures for accurate refitting. Our approach employs a Bayesian hierarchical model with a sparsity-inducing Laplace prior, and the Shotgun Stochastic Search (SSS) algorithm to efficiently explore possible signature subsets and identify the optimal one. We established the model selection consistency of BayesSigRefitting and demonstrated, through simulation and real data studies across seven cancer types, that it outperformed existing methods in both signature selection and signature activity estimation. These findings highlight the potential of BayesSigRefitting to enhance the accuracy and reliability of mutational signature analysis, especially in settings with limited sample sizes.

REFERENCES

ALEXANDROV, L. B., JU, Y. S., HAASE, K., VAN LOO, P., MARTINCORENA, I., NIK-ZAINAL, S., TOTOKI, Y., FUJIMOTO, A., NAKAGAWA, H. et al. (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science* **354** 618–622.

ALEXANDROV, L. B., KIM, J., HARADHVALA, N. J., HUANG, M. N., TIAN NG, A. W., WU, Y., BOOT, A., COVINGTON, K. R., GORDENIN, D. A. et al. (2020). The repertoire of mutational signatures in human cancer. *Nature* **578** 94–101.

ALEXANDROV, L. B., NIK-ZAINAL, S., WEDGE, D. C., CAMPBELL, P. J. and STRATTON, M. R. (2013). Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*. **3** 246–259.

ALEXANDROV, L. B. and STRATTON, M. R. (2014). Mutational signatures: The patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24** 52–60.

BAEZ-ORTEGA, A. and GORI, K. (2019). Computational approaches for discovery of mutational signatures in cancer. *Brief. Bioinform*. **20** 77–88.

BLOKZIJL, F., JANSSEN, R., VAN BOXTEL, R. and CUPPEN, E. (2018). MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. *Gen. Med*. **10** 1–11.

CARTOLANO, M., ABEDPOUR, N., ACHTER, V., YANG, T.-P., ACKERMANN, S., FISCHER, M. and PEIFER, M. (2020). CaMuS: Simultaneous fitting and de novo imputation of cancer mutational signature. *Sci. Rep*. **10** 19316.

Key words and phrases. Bayesian model selection, cancer genomics, mutational signatures, signature refitting, somatic mutations.

CEMGIL, A. T. (2009). Bayesian inference for nonnegative matrix factorisation models. *Comput. Intell. Neurosci.* **2009** 785152.

DIAZ-GAY, M., VANGARA, R., BARNES, M., WANG, X., ISLAM, S. A., VERMES, I., NARASIMMAN, N. B., YANG, T., JIANG, Z. et al. (2023). Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. bioRxiv 2023-07.

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

HANS, C., DOBRA, A. and WEST, M. (2007). Shotgun stochastic search for "large p" regression. *J. Amer. Statist. Assoc.* **102** 507–516. MR2370849

HUA, M. and ZHU, B. (2026). Supplement to "Bayesian selection and refitting of reference mutational signatures in cancer genomics." https://doi.org/10.1214/26-AOAS2138SUPPA, https://doi.org/10.1214/26-AOAS2138SUPPB, https://doi.org/10.1214/26-AOAS2138SUPPC, https://doi.org/10.1214/26-AOAS2138SUPPD https://doi.org/10.1214/26-AOAS2138SUPPE, https://doi.org/10.1214/26-AOAS2138SUPPF

HUANG, X., WOJTOWICZ, D. and PRZYTYCKA, T. M. (2018). Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics* **34** 330–337.

HÜBSCHMANN, D., JOPP-SAILE, L., ANDRESEN, C., KRÄMER, S., GU, Z., HEILIG, C. E., KREUTZFELDT, S., TELEANU, V., FRÖHLING, S. et al. (2021). Analysis of mutational signatures with yet another package for signature analysis. *Genes Chromosomes Cancer* **60** 314–331.

KUCAB, J. E., ZOU, X., MORGANELLA, S., JOEL, M., NANDA, A. S., NAGY, E., GOMEZ, C., DEGASPERI, A., HARRIS, R. et al. (2019). A compendium of mutational signatures of environmental agents. *Cell* **177** 821–836.

LEE, D., WANG, D., YANG, X. R., SHI, J., LANDI, M. T. and ZHU, B. (2022). SUITOR: Selecting the number of mutational signatures through cross-validation. *PLoS Comput. Biol.* **18** e1009309.

LEE, D. D. and SEUNG, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401** 788–791.

LEE, J., LEE, A. J., LEE, J.-K., PARK, J., KWON, Y., PARK, S., CHUN, H., JU, Y. S. and HONG, D. (2018). Mutalisk: A web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. *Nucleic Acids Res.* **46** W102–W108.

LI, S., CRAWFORD, F. W. and GERSTEIN, M. B. (2020). Using sigLASSO to optimize cancer mutation signatures jointly with sampling likelihood. *Nat. Commun.* **11** 3575.

LYNCH, A. G. (2016). Decomposition of mutational context signatures using quadratic programming methods. *F1000Res.* **5** 1253.

MAURA, F., DEGASPERI, A., NADEU, F., LEONGAMORNLERT, D., DAVIES, H., MOORE, L., ROYO, R., ZICCHEDDU, B., PUENTE, X. et al. (2019). A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* **10** 2969.

NIK-ZAINAL, S., DAVIES, H., STAAF, J., RAMAKRISHNA, M., GLODZIK, D., ZOU, X., MARTINCORENA, I., ALEXANDROV, L. B., MARTIN, S. et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534** 47–54.

OMICHESSAN, H., SEVERI, G. and PERDUCA, V. (2019). Computational tools to detect signatures of mutational processes in DNA from tumours: A review and empirical comparison of performance. *PLoS ONE* **14** e0221235.

PETLJAK, M., ALEXANDROV, L. B., BRAMMELD, J. S., PRICE, S., WEDGE, D. C., GROSSMANN, S., DAWSON, K. J., JU, Y. S., IORIO, F. et al. (2019). Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176** 1282–1294.

PFEIFER, G. P. (2010). Environmental exposures and mutational patterns of cancer genomes. *Gen. Med.* **2** 1–4.

POLAK, P., KIM, J., BRAUNSTEIN, L. Z., KARLIC, R., HARADHAVALA, N. J., TIAO, G., ROSEBROCK, D., LIVITZ, D., KÜBLER, K. et al. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **49** 1476–1486.

ROSALES, R. A., DRUMMOND, R. D., VALIERIS, R., DIAS-NETO, E. and DA SILVA, I. T. (2017). signeR: An empirical Bayesian approach to mutational signature discovery. *Bioinformatics* **33** 8–16.

ROSENTHAL, R., MCGRANAHAN, N., HERRERO, J., TAYLOR, B. S. and SWANTON, C. (2016). DeconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17** 1–11.

RUSTAD, E. H., NADEU, F., ANGELOPOULOS, N., ZICCHEDDU, B., BOLLI, N., PUENTE, X. S., CAMPO, E., LANDGREN, O. and MAURA, F. (2021). mmsig: A fitting approach to accurately identify somatic mutational signatures in hematological malignancies. *Commun. Biol.* **4** 424.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1078–1098
https://doi.org/10.1214/26-AOAS2179

GMI: GROUP-LEVEL MAIN EFFECTS AND INTERACTIONS IN HIGH-DIMENSIONAL DATA WITH APPLICATIONS TO PATHWAY AND INTERACTION DISCOVERY IN GENE EXPRESSION ANALYSIS

BY JINYU NIE[1,a], LI LIU[2,c], TAOBO HU[3,d], WEI LIU[4,e] AND HUAZHEN LIN[1,b]

[1]*Center of Statistical Research and School of Statistics, New Cornerstone Science Laboratory, Southwestern University of Finance and Economics,* [a]*17713543862@163.com,* [b]*linhz@swufe.edu.cn*

[2]*School of Mathematics and Statistics, Wuhan University,* [c]*lliu.math@whu.edu.cn*

[3]*Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University,* [d]*Taobo.hu@scilifelab.se*

[4]*School of Mathematics, Sichuan University,* [e]*liuwei8@scu.edu.cn*

Genetic interactions are essential for understanding the risk and progression of complex diseases. However, signals from individual genes and their pairwise interactions are often weak; most phenotypes are driven by alterations in a limited number of pathways and interactions between them. Identifying such pathways and their interactions is critical in biomedical research. Although traditional analyses have extended beyond main effects to include gene-gene interactions, most existing methods remain at the gene level and fail to capture higher-level pathway interactions. In this paper we propose a novel group-level model that jointly identifies key pathways and their interactions associated with clinical outcomes such as disease status or survival. The model involves estimating a high-dimensional binary matrix, which presents significant computational challenges. To overcome this, we reformulate the problem as a standard high-dimensional estimation task with hierarchical and exclusivity constraints and develop a two-stage estimation procedure. Theoretical analysis, simulation studies, and applications to TCGA breast cancer and Michigan lung cancer datasets demonstrate the superior performance of our method. In particular, our approach yields biologically meaningful insights, reveals novel gene-pathway mechanisms, and achieves substantially improved prediction accuracy and sensitivity, with comparable specificity to competing methods, including those modeling gene-gene interactions or employing two-step procedures that separately estimate pathways and their interactions.

REFERENCES

BABCHIA, N., CALIPEL, A., MOURIAUX, F., FAUSSAT, A.-M. and MASCARELLI, F. (2010). The PI3K/Akt and mTOR/P70S6K signaling pathways in human uveal melanoma cells: Interaction with B-Raf/ERK. *Investig. Ophthalmol. Vis. Sci.* **51** 421–429. https://doi.org/10.1167/iovs.09-3974

BACHELDER, R. E., LIPSCOMB, E. A., LIN, X., WENDT, M. A., CHADBORN, N. H., EICKHOLT, B. J. and MERCURIO, A. M. (2003). Competing autocrine pathways involving alternative neuropilin-1 ligands regulate Chemotaxis of carcinoma cells. *Cancer Res.* **63** 5230–5233.

BEER, D. G., KARDIA, S. L., HUANG, C.-C., GIORDANO, T. J., LEVIN, A. M., MISEK, D. E., LIN, L., CHEN, G., GHARIB, T. G. et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* **8** 816–824. https://doi.org/10.1038/nm733

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

BONDELL, H. D. and REICH, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64** 115–123, 322–323. MR2422825 https://doi.org/10.1111/j.1541-0420.2007.00843.x

BOURAS, T., PAL, B., VAILLANT, F., HARBURG, G., ASSELIN-LABAT, M.-L., OAKES, S. R., LINDEMAN, G. J. and VISVADER, J. E. (2008). Notch signaling regulates mammary stem cell function and luminal cell-fate commitment. *Cell Stem Cell* **3** 429–441. https://doi.org/10.1016/j.stem.2008.08.001

Key words and phrases. High dimension, higher-level interaction analysis, variable selection.

CHAKRABARTI, R., CELIÀ-TERRASSA, T., KUMAR, S., HANG, X., WEI, Y., CHOUDHURY, A., HWANG, J., PENG, J., NIXON, B. et al. (2018). Notch ligand Dll1 mediates cross-talk between mammary stem cells and the macrophageal niche. *Science* **360** eaan4153. https://doi.org/10.1126/science.aan4153

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

CHIPMAN, H., HAMADA, M. and WU, C. (1997). A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics* **39** 372–381. https://doi.org/10.1080/00401706.1997.10485156

CHITNIS, M. M., YUEN, J. S., PROTHEROE, A. S., POLLAK, M. and MACAULAY, V. M. (2008). The type 1 insulin-like growth factor receptor pathway. *Clin. Cancer Res.* **14** 6364–6370. https://doi.org/10.1158/1078-0432.CCR-07-4879

CHOI, N. H., LI, W. and ZHU, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *J. Amer. Statist. Assoc.* **105** 354–364. MR2656056 https://doi.org/10.1198/jasa.2010.tm08281

CLEVENGER, C. V., FURTH, P. A., HANKINSON, S. E. and SCHULER, L. A. (2003). The role of prolactin in mammary carcinoma. *Endocr. Rev.* **24** 1–27. https://doi.org/10.1210/er.2001-0036

CORDELL, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.* **10** 392–404. https://doi.org/10.1038/nrg2579

CREIGHTON, C. J., CASA, A., LAZARD, Z., HUANG, S., TSIMELZON, A., HILSENBECK, S. G., OSBORNE, C. K. and LEE, A. V. (2008). Insulin-like growth factor-I activates gene transcription programs strongly associated with poor breast cancer prognosis. *J. Clin. Oncol.* **26** 4078–4085. https://doi.org/10.1200/JCO.2007.13.4429

ELHAMAMSY, A. R., METGE, B. J., ALSHEIKH, H. A., SHEVDE, L. A. and SAMANT, R. S. (2022). Ribosome biogenesis: A central player in cancer metastasis and therapeutic resistance. *Cancer Res.* **82** 2344–2353. https://doi.org/10.1158/0008-5472.CAN-21-4087

FAN, Y., KONG, Y., LI, D. and ZHENG, Z. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *Ann. Statist.* **43** 1243–1272. MR3346702 https://doi.org/10.1214/14-AOS1308

GOEL, H. L. and MERCURIO, A. M. (2013). VEGF targets the tumour cell. *Nat. Rev. Cancer* **13** 871–882. https://doi.org/10.1038/nrc3627

GUJRAL, T. S., CHAN, M., PESHKIN, L., SORGER, P. K., KIRSCHNER, M. W. and MACBEATH, G. (2014). A noncanonical Frizzled2 pathway regulates epithelial-mesenchymal transition and metastasis. *Cell* **159** 844–856. https://doi.org/10.1016/j.cell.2014.10.032

HAMADA, M. and WU, C. J. (1992). Analysis of designed experiments with complex aliasing. *J. Qual. Technol.* **24** 130–137. https://doi.org/10.1080/00224065.1992.11979383

HANKINSON, S. E., COLDITZ, G. A. and WILLETT, W. C. (2004). Towards an integrated model for breast cancer etiology: The lifelong interplay of genes, lifestyle, and hormones. *Breast Cancer Res.* **6** 213.

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

HAO, N. and ZHANG, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **109** 1285–1301. MR3265697 https://doi.org/10.1080/01621459.2014.881741

HE, Y., ZHOU, L., XIA, Y. and LIN, H. (2023). Center-augmented ℓ_2-type regularization for subgroup learning. *Biometrics* **79** 2157–2170. MR4643984 https://doi.org/10.1111/biom.13725

HERBST, R. S., MORGENSZTERN, D. and BOSHOFF, C. (2018). The biology and management of non-small cell lung cancer. *Nature* **553** 446–454.

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218. https://doi.org/10.1007/BF01908075

KAVARTHAPU, R., ANBAZHAGAN, R. and DUFAU, M. L. (2021). Crosstalk between PRLR and EGFR/HER2 signaling pathways in breast cancer. *Cancers* **13** 4685.

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

KIM, J., LIM, J., KIM, Y. and JANG, W. (2018). Bayesian variable selection with strong heredity constraints. *J. Korean Statist. Soc.* **47** 314–329. MR3840864 https://doi.org/10.1016/j.jkss.2018.03.003

KIM, S. H., CHEN, G., KING, A. N., JEON, C. K., CHRISTENSEN, P. J., ZHAO, L., SIMPSON, R. U., THOMAS, D. G., GIORDANO, T. J. et al. (2012). Characterization of vitamin D receptor (VDR) in lung adenocarcinoma. *Lung Cancer* **77** 265–271. https://doi.org/10.1016/j.lungcan.2012.04.010

KONG, Y., LI, D., FAN, Y. and LV, J. (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *Ann. Statist.* **45** 897–922. MR3650404 https://doi.org/10.1214/16-AOS1474

KOOPERBERG, C. and LEBLANC, M. (2008). Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet. Epidemiol.* **32** 255. https://doi.org/10.1002/gepi.20300

LAWAL, A. K., ROTTER, T., KINSMAN, L., MACHOTTA, A., RONELLENFITSCH, U., SCOTT, S. D., GOODRIDGE, D., PLISHKA, C. and GROOT, G. (2016). What is a clinical pathway? Refinement of an op-

erational definition to identify clinical pathway studies for a Cochrane systematic review. *BMC Med*. **14** 1–5. https://doi.org/10.1186/s12916-016-0580-z

LI, D., KONG, Y., FAN, Y. and LV, J. (2022). High-dimensional interaction detection with false sign rate control. *J. Bus. Econom. Statist*. **40** 1234–1245. MR4439285 https://doi.org/10.1080/07350015.2021.1917419

LI, Y. and LIU, J. S. (2019). Robust variable and interaction selection for logistic regression and general index models. *J. Amer. Statist. Assoc*. **114** 271–286. MR3941254 https://doi.org/10.1080/01621459.2017.1401541

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

LIU, C., MA, J. and AMOS, C. I. (2015). Bayesian variable selection for hierarchical gene-environment and gene-gene interactions. *Hum. Genet*. **134** 23–36. https://doi.org/10.1007/s00439-014-1478-5

LOVE, M. I., HUBER, W. and ANDERS, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. **15** 550. https://doi.org/10.1186/s13059-014-0550-8

NELDER, J. A. (1977). A reformulation of linear models. *J. Roy. Statist. Soc. Ser. A* **140** 48–76. MR0458743 https://doi.org/10.2307/2344517

NIE, J., LIU, L., HU, T., LIU, W. and LIN, H. (2026). Supplement to "Gmi: Group-level main effects and interactions in high-dimensional data with applications to pathway and interaction discovery in gene expression analysis." https://doi.org/10.1214/26-AOAS2179SUPPA, https://doi.org/10.1214/26-AOAS2179SUPPB

NOLAN, E., LINDEMAN, G. J. and VISVADER, J. E. (2023). Deciphering breast cancer: From biology to the clinic. *Cell* **186** 1708–1728. https://doi.org/10.1016/j.cell.2023.01.040

SANCHEZ-VEGA, F., MINA, M., ARMENIA, J., CHATILA, W. K., LUNA, A., LA, K. C., DIMITRIADOY, S., LIU, D. L., KANTHETI, H. S. et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell* **173** 321–337. https://doi.org/10.1016/j.cell.2018.03.035

SHAH, R. D. (2016). Modelling interactions in high-dimensional data with backtracking. *J. Mach. Learn. Res*. **17** Paper No. 207, 31. MR3595141

SHARMA, P., ALSHARIF, S., FALLATAH, A. and CHUNG, B. M. (2019). Intermediate filaments as effectors of cancer development and metastasis: A focus on keratins, vimentin, and nestin. *Cells* **8** 497. https://doi.org/10.3390/cells8050497

SHE, Y., WANG, Z. and JIANG, H. (2018). Group regularized estimation under structural hierarchy. *J. Amer. Statist. Assoc*. **113** 445–454. MR3803477 https://doi.org/10.1080/01621459.2016.1260470

SHEN, X. and HUANG, H.-C. (2010). Grouping pursuit through a regularization solution surface. *J. Amer. Statist. Assoc*. **105** 727–739. MR2724856 https://doi.org/10.1198/jasa.2010.tm09380

SHI, L., QIU, F., SHI, C., ZHANG, G. and YU, F. (2025). Integrative bulk RNA analysis unveils immune evasion mechanisms and predictive biomarkers of osimertinib resistance in non-small cell lung cancer. *Discov. Oncol*. **16** 1541.

STECCA, B., MAS, C., CLEMENT, V., ZBINDEN, M., CORREA, R., PIGUET, V., BEERMANN, F. and RUIZ I ALTABA, A. (2007). Melanomas require HEDGEHOG-GLI signaling regulated by interactions between GLI1 and the RAS-MEK/AKT pathways. *Proc. Natl. Acad. Sci. USA* **104** 5895–5900. https://doi.org/10.1073/pnas.0700776104

TANG, C. Y., FANG, E. X. and DONG, Y. (2020). High-dimensional interactions detection with sparse principal Hessian matrix. *J. Mach. Learn. Res*. **21** Paper No. 19, 25. MR4071202

TAYLOR, K. M., MORGAN, H. E., SMART, K., ZAHARI, N. M., PUMFORD, S., ELLIS, I. O., ROBERTSON, J. F. and NICHOLSON, R. I. (2007). The emerging role of the LIV-1 subfamily of zinc transporters in breast cancer. *Mol. Med*. **13** 396–406. https://doi.org/10.2119/2007-00040

THANEI, G.-A., MEINSHAUSEN, N. and SHAH, R. D. (2018). The xyz algorithm for fast interaction search in high-dimensional data. *J. Mach. Learn. Res*. **19** Paper No. 37, 42. MR3862444

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

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

XIE, H., HANAI, J.-I., REN, J.-G., KATS, L., BURGESS, K., BHARGAVA, P., SIGNORETTI, S., BILLIARD, J., DUFFY, K. J. et al. (2014). Targeting lactate dehydrogenase-a inhibits tumorigenesis and tumor progression in mouse models of lung cancer and impacts tumor-initiating cells. *Cell Metab*. **19** 795–809. https://doi.org/10.1016/j.cmet.2014.03.003

YE, G.-B. and XIE, X. (2011). Split Bregman method for large scale fused Lasso. *Comput. Statist. Data Anal*. **55** 1552–1569. MR2748661 https://doi.org/10.1016/j.csda.2010.10.021

YUAN, M., JOSEPH, V. R. and ZOU, H. (2009). Structured variable selection and estimation. *Ann. Appl. Stat*. **3** 1738–1757. MR2752156 https://doi.org/10.1214/09-AOAS254

ZHANG, C., DU, Y., JI, Y., YE, X., LIAN, J., ZHOU, H., GAO, Z., XU, H., TANG, Y. et al. (2025). Lactylation-driven KRT19 promotes non-small cell lung cancer progression by suppressing cellular senescence. *J. Exp. Clin. Cancer Res*. https://doi.org/10.1186/s13046-025-03602-5

ZHANG, Y., MOERKENS, M., RAMAIAHGARI, S., DE BONT, H., PRICE, L., MEERMAN, J. and VAN DE WATER, B. (2011). Elevated insulin-like growth factor 1 receptor signaling induces antiestrogen resistance through the MAPK/ERK and PI3K/Akt signaling routes. *Breast Cancer Res*. **13** R52.

ZHOU, M., DAI, M., YAO, Y., LIU, J., YANG, C. and PENG, H. (2023). BOLT-SSI: A statistical approach to screening interaction effects for ultra-high dimensional data. *Statist. Sinica* **33** 2327–2358. MR4647037 https://doi.org/10.1007/jhep04(2023)080

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1099–1123
https://doi.org/10.1214/26-AOAS2196

ROBUST HIGH-THROUGHPUT IMAGING ANALYSIS WITH WASSERSTEIN GEODESIC TRANSFORMATIONS

BY GREGORY J. HUNT[1,a] AND JOHANN A. GAGNON-BARTSCH[2,b]

[1]*Department of Mathematics, William & Mary,* [a]*ghunt@wm.edu*
[2]*Department of Statistics, University of Michigan,* [b]*johanngb@umich.edu*

High-throughput cell imaging has been increasingly used in drug discovery to simultaneously profile the morphological response of cells to thousands of compounds using high-resolution microscopy and automated image analysis. Such experiments characterize thousands of image features in millions of cells across thousands of experimental conditions. Analytical difficulties arise with this scale of analysis as many features have distributions with extremely long tails, high skewness, remote outliers, and high-leverage points. This makes important signals difficult to find and means analyses are often sensitive to individual observations or features.

This work considers a recent high-quality Cell Painting dataset profiling compounds from the EU-OPENSCREEN consortium. The study perturbs HepG2 human liver cancer cells in order to morphologically profile cellular response to the compounds. Without adjustment, analysis of the imaging data is hampered by long-tailed distributions and outliers. To combat this, we introduce Wasserstein Geodesic Transformations (WGTs), a new approach that adaptively moves features in Wasserstein space to make downstream analysis less ad hoc, more stable, and more scalable.

In application to the Cell Painting data, WGTs substantially improve data analysis by enhancing visualization, improving compound clustering, and stabilizing analyses. They also help uncover unwanted spatial effects arising from plate layout, explaining some outlying compound responses. Overall, WGTs achieve performance comparable with more aggressive transformations, while inducing less distortion, artifacts, and better preserving distributional features. More broadly, the adaptivity of WGT makes it promising for a wide range of cell imaging pipelines.

REFERENCES

BICKEL, P. J. and DOKSUM, K. A. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.* **76** 296–311. MR0624332

BOLSTAD, B. M., IRIZARRY, R. A., ÅSTRAND, M. and SPEED, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19** 185–193.

BOUGEN-ZHUKOV, N., LOH, S. Y., LEE, H. K. and LOO, L.-H. (2017). Large-scale image-based screening and profiling of cellular phenotypes. *Cytometry, Part A* **91** 115–125.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. (With discussion). *J. Roy. Statist. Soc. Ser. B, Methodol.* **26** 211–252. MR0192611

CAICEDO, J. C., COOPER, S., HEIGWER, F., WARCHAL, S., QIU, P., MOLNAR, C., VASILEVICH, A. S., BARRY, J. D., BANSAL, H. S. et al. (2017). Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14** 849–863.

CAICEDO, J. C., SINGH, S. and CARPENTER, A. E. (2016). Applications in image-based profiling of perturbations. *Curr. Opin. Biotechnol.* **39** 134–142.

CHANDRASEKARAN, S. N., CEULEMANS, H., BOYD, J. D. and CARPENTER, A. E. (2021). Image-based profiling for drug discovery: Due for a machine-learning upgrade? *Nat. Rev. Drug Discov.* **20** 145–159.

DESGRAUPES, B. (2023). clusterCrit: Clustering Indices R package version 1.3.0.

HARRISON, P. J., GUPTA, A., RIETDIJK, J., WIESLANDER, H., CARRERAS-PUIGVERT, J., GEORGIEV, P., WÄHLBY, C., SPJUTH, O. and SINTORN, I.-M. (2023). Evaluating the utility of brightfield image data for mechanism of action prediction. *PLoS Comput. Biol.* **19** e1011323.

Key words and phrases. Image-based drug profiling, transformation.

HUNT, G. J., DANE, M. A., KORKOLA, J. E., HEISER, L. M. and GAGNON-BARTSCH, J. A. (2020). Automatic transformation and integration to improve visualization and discovery of latent effects in imaging data. *J. Comput. Graph. Statist.* **29** 929–941. MR4191252 https://doi.org/10.1080/10618600.2020.1741379

HUNT, G. J and GAGNON-BARTSCH, J. A (2026). Supplement to "Robust high-throughput imaging analysis with Wasserstein geodesic transformations." https://doi.org/10.1214/26-AOAS2196SUPPA, https://doi.org/10.1214/26-AOAS2196SUPPB

KAMENTSKY, L., JONES, T. R., FRASER, A., BRAY, M.-A., LOGAN, D. J., MADDEN, K. L., LJOSA, V., RUEDEN, C., ELICEIRI, K. W. et al. (2011). Improved structure, function and compatibility for CellProfiler: Modular high-throughput image analysis software. *Bioinformatics* **27** 1179–1180. https://doi.org/10.1093/bioinformatics/btr095

LEE, J.-Y. J., MILLER, J. A., BASU, S., KEE, T.-Z. V. and LOO, L.-H. (2018). Building predictive in vitro pulmonary toxicity assays using high-throughput imaging and artificial intelligence. *Arch. Toxicol.* **92** 2055–2075.

PANARETOS, V. M. and ZEMEL, Y. (2020). *An Invitation to Statistics in Wasserstein Space*. *SpringerBriefs in Probability and Mathematical Statistics*. Springer, Cham. MR4350694 https://doi.org/10.1007/978-3-030-38438-8

ROHBAN, M. H., SINGH, S., WU, X., BERTHET, J. B., BRAY, M.-A., SHRESTHA, Y., VARELAS, X., BOEHM, J. S. and CARPENTER, A. E. (2017). Systematic morphological profiling of human gene and allele function via Cell Painting. *eLife* **6**.

SANTAMBROGIO, F. (2015). *Optimal Transport for Applied Mathematicians*: *Calculus of Variations*, *PDEs*, *and Modeling*. *Progress in Nonlinear Differential Equations and Their Applications* **87**. Birkhäuser/Springer, Cham. MR3409718 https://doi.org/10.1007/978-3-319-20828-2

SCHEEDER, C., HEIGWER, F. and BOUTROS, M. (2018). Machine learning and image-based profiling in drug discovery. *Curr. Opin. Syst. Biol.* **10** 43–52. https://doi.org/10.1016/j.coisb.2018.05.004

SIMM, J., KLAMBAUER, G., ARANY, A., STEIJAERT, M., WEGNER, J. K., GUSTIN, E., CHUPAKHIN, V., CHONG, Y. T., VIALARD, J. et al. (2018). Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell Chem. Biol.* **25** 611–618.e3. https://doi.org/10.1016/j.chembiol.2018.01.015

SINGH, S., CARPENTER, A. E. and GENOVESIO, A. (2014). Increasing the content of high-content screening: An overview. *J. Biomol. Screen.* **19** 640–650.

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

WILLIS, C., NYFFELER, J. and HARRILL, J. (2020). Phenotypic profiling of reference chemicals across biologically diverse cell types using the Cell Painting assay. *SLAS Discov.* **25** 755–769.

WOLFF, C., NEUENSCHWANDER, M., BEESE, C. J., SITANI, D., RAMOS, M. C., SROVNALOVA, A., VARELA, M. J., POLISHCHUK, P., SKOPELITOU, K. E. et al. (2025). Morphological profiling data resource enables prediction of chemical compound properties. *iScience* **28** 112445. https://doi.org/10.1016/j.isci.2025.112445

YOUNG, D. W., BENDER, A., HOYT, J., MCWHINNIE, E., CHIRN, G.-W., TAO, C. Y., TALLARICO, J. A., LABOW, M., JENKINS, J. L. et al. (2008). Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.* **4** 59–68.

ZANELLA, F., LORENS, J. B. and LINK, W. (2010). High content screening: Seeing is believing. *Trends Biotechnol.* **28** 237–245.

ZIEGLER, S., SIEVERS, S. and WALDMANN, H. (2021). Morphological profiling of small molecules. *Cell Chem. Biol.* **28** 300–319.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1124–1147
https://doi.org/10.1214/26-AOAS2149

3D BIVARIATE SPATIAL MODELLING OF ARGO OCEAN TEMPERATURE AND SALINITY

BY MARY LAI O. SALVAÑA[1,a], JIAN CAO[2,b] AND MIKYOUNG JUN[2,c]

[1]*Department of Statistics, University of Connecticut,* [a]*marylai.salvana@uconn.edu*
[2]*Department of Mathematics, University of Houston,* [b]*jcao21@central.uh.edu,* [c]*mjun@central.uh.edu*

Variables within the global oceans can reveal the impacts of a warming climate, as the oceans absorb huge amounts of solar energy. Understanding the joint spatial distribution of key ocean variables is, therefore, essential. In this paper we investigate the spatial dependence structure between ocean temperature and salinity using Argo observations and construct a bivariate spatial model covering from the surface through the ocean interior. We develop a flexible class of multivariate nonstationary covariance models defined in three-dimensional (3D) space (longitude × latitude × depth) that allow the variances and correlations to vary with depth, capturing the ocean's vertical structure. These models describe the joint spatial distribution of the two variables while incorporating the underlying vertical structure of the ocean. We apply this framework to Argo temperature and salinity data and address the computational challenges of large data volumes through the Vecchia approximation. Our results show that the proposed bivariate covariance model effectively represents the complex vertical cross-covariance structure of the processes and their first- and second-order differences, whereas classical bivariate models, including the bivariate Matérn, poorly fit the empirical cross-covariance structure.

REFERENCES

AL SENAFI, F. and ANIS, A. (2020). Internal waves on the continental shelf of the northwestern Arabian Gulf. *Front. Mar. Sci.* **805**.

ALEGRÍA, A. (2020). Cross-dimple in the cross-covariance functions of bivariate isotropic random fields on spheres. *Stat* **9** e301. MR4156481 https://doi.org/10.1002/sta4.301

BÖHME, L. and SEND, U. (2005). Objective analyses of hydrographic data for referencing profiling float salinities in highly variable environments. *Deep-Sea Res., Part* 2, *Top. Stud. Oceanogr.* **52** 651–664.

CABANES, C., THIERRY, V. and LAGADEC, C. (2016). Improvement of bias detection in Argo float conductivity sensors and its application in the North Atlantic. *Deep-Sea Res., Part* 1, *Oceanogr. Res. Pap.* **114** 128–136.

CAO, J., GUINNESS, J., GENTON, M. G. and KATZFUSS, M. (2022). Scalable Gaussian-process regression and variable selection using Vecchia approximations. *J. Mach. Learn. Res.* **23** 348. MR4577787

CAO, J., ZHANG, J., SUN, Z. and KATZFUSS, M. (2024). Locally anisotropic nonstationary covariance functions on the sphere. *J. Agric. Biol. Environ. Stat.* **29** 212–231. MR4741604 https://doi.org/10.1007/s13253-023-00573-y

CHAIGNEAU, A., LE TEXIER, M., ELDIN, G., GRADOS, C. and PIZARRO, O. (2011). Vertical structure of mesoscale eddies in the eastern South Pacific Ocean: A composite analysis from altimetry and Argo profiling floats. *J. Geophys. Res., Oceans* **116**.

CHEN, G. and GENG, D. (2019). A "mirror layer" of temperature and salinity in the ocean. *Clim. Dyn.* **52** 1–13.

CHEN, G., PENG, L. and MA, C. (2018). Climatology and seasonality of upper ocean salinity: A three-dimensional view from Argo floats. *Clim. Dyn.* **50** 2169–2182.

CHEN, G. and WANG, X. (2016). Vertical structure of upper-ocean seasonality: Annual and semiannual cycles with oceanographic implications. *J. Climate* **29** 37–59.

CHEN, L., ZHANG, R.-H. and GAO, C. (2022). Effects of temperature and salinity on surface currents in the equatorial Pacific. *J. Geophys. Res., Oceans* **127**. e2021JC018175.

DING, D.-S., PATEL, A. K., SINGHANIA, R. R., CHEN, C.-W. and DONG, C.-D. (2022). Effects of temperature and salinity on growth, metabolism and digestive enzymes synthesis of goniopora columna. *Biology* **11** 436.

Key words and phrases. 3D covariance functions, Argo, cross-covariance function, nonstationary, salinity, spatial, temperature.

DONG, C., MCWILLIAMS, J. C., LIU, Y. and CHEN, D. (2014). Global heat and salt transports by eddy movement. *Nat. Commun.* **5** 1–6.

ESCOBAR, A., NEGRO, V., LÓPEZ-GUTIÉRREZ, J. S. and ESTEBAN, M. (2016). Influence of temperature and salinity on hydrodynamic forces. *J. Ocean Eng. Sci.* **1** 325–336.

GALÁN, A., SALDÍAS, G. S., CORREDOR-ACOSTA, A., MUÑOZ, R., LARA, C. and IRIARTE, J. L. (2021). Argo float reveals biogeochemical characteristics along the freshwater gradient off western Patagonia. *Front. Mar. Sci.* **8** 613265.

GANGOPADHYAY, A. (2022). *Introduction to Ocean Circulation and Modeling*. CRC Press, Boca Raton.

GNEITING, T., KLEIBER, W. and SCHLATHER, M. (2010). Matérn cross-covariance functions for multivariate random fields. *J. Amer. Statist. Assoc.* **105** 1167–1177. MR2752612 https://doi.org/10.1198/jasa.2010.tm09420

GOOD, S. A., MARTIN, M. J. and RAYNER, N. A. (2013). EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *J. Geophys. Res., Oceans* **118** 6704–6716.

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

HELBER, R. W., KARA, A. B., RICHMAN, J. G., CARNES, M. R., BARRON, C. N., HURLBURT, H. E. and BOYER, T. (2012). Temperature versus salinity gradients below the ocean mixed layer. *J. Geophys. Res., Oceans* **117**.

HOSODA, S., OHIRA, T. and NAKAMURA, T. (2008). A monthly mean dataset of global Oceanic temperature and salinity derived from Argo float observations. *JAMSTEC Rep. Res. Dev.* **8** 47–59.

HU, A. J., KUUSELA, M., LEE, A. B., GIGLIO, D. and WOOD, K. M. (2024). Spatiotemporal methods for estimating subsurface ocean thermal response to tropical cyclones. *Adv. Stat. Climatol. Meteorol. Oceanogr.* **10** 69–93.

JANA, S., GANGOPADHYAY, A., HALEY, P. J. and LERMUSIAUX, P. F. (2022). Sound speed variability over Bay of Bengal from Argo observations (2011-2020). In *OCEANS* 2022-*Chennai* 1–8. IEEE.

JEONG, J., JUN, M. and GENTON, M. G. (2017). Spherical process models for global spatial statistics. *Statist. Sci.* **32** 501–513. MR3730519 https://doi.org/10.1214/17-STS620

JOHNSON, G. C., HOSODA, S., JAYNE, S. R., OKE, P. R., RISER, S. C., ROEMMICH, D., SUGA, T., THIERRY, V., WIJFFELS, S. E. et al. (2022). Argo-two decades: Global oceanography, revolutionized. *Annu. Rev. Mar. Sci.* **14** 379–403.

JUN, M. (2011). Non-stationary cross-covariance models for multivariate processes on a globe. *Scand. J. Stat.* **38** 726–747. MR2859747 https://doi.org/10.1111/j.1467-9469.2011.00751.x

JUN, M. and STEIN, M. L. (2007). An approach to producing space-time covariance functions on spheres. *Technometrics* **49** 468–479. MR2394558 https://doi.org/10.1198/004017007000000155

JUN, M. and STEIN, M. L. (2008). Nonstationary covariance models for global data. *Ann. Appl. Stat.* **2** 1271–1289. MR2655659 https://doi.org/10.1214/08-AOAS183

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

KLEIBER, W. and GENTON, M. G. (2013). Spatially varying cross-correlation coefficients in the presence of nugget effects. *Biometrika* **100** 213–220. MR3034334 https://doi.org/10.1093/biomet/ass057

KLEIBER, W. and NYCHKA, D. (2012). Nonstationary modeling for multivariate spatial processes. *J. Multivariate Anal.* **112** 76–91. MR2957287 https://doi.org/10.1016/j.jmva.2012.05.011

KUNOTH, A., LYCHE, T., SANGALLI, G., SERRA-CAPIZZANO, S., MANNI, C. and SPELEERS, H. (2018). *Splines and PDEs*: *From approximation theory to numerical linear algebra*. Springer, Cham.

KUUSELA, M. and STEIN, M. L. (2018). Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proc. R. Soc. A* **474** 20180400.

LI, H., XU, F., ZHOU, W., WANG, D., WRIGHT, J. S., LIU, Z. and LIN, Y. (2017). Development of a global gridded Argo data set with Barnes successive corrections. *J. Geophys. Res., Oceans* **122** 866–889.

LIU, C., LIANG, X., CHAMBERS, D. P. and PONTE, R. M. (2020a). Global patterns of spatial and temporal variability in salinity from multiple gridded Argo products. *J. Climate* **33** 8751–8766.

LIU, H., ONG, Y.-S., SHEN, X. and CAI, J. (2020b). When Gaussian process meets big data: A review of scalable GPs. *IEEE Trans. Neural Netw. Learn. Syst.* **31** 4405–4423. MR4169962 https://doi.org/10.1109/tnnls.2019.2957109

MAES, C. and O'KANE, T. J. (2014). Seasonal variations of the upper ocean salinity stratification in the Tropics. *J. Geophys. Res., Oceans* **119** 1706–1722.

MCPHADEN, M. J. and HAYES, S. P. (1991). On the variability of winds, sea surface temperature, and surface layer heat content in the western equatorial Pacific. *J. Geophys. Res., Oceans* **96** 3331–3342.

MERCHEL, M., WALCZOWSKI, W., RAK, D. and WIECZOREK, P. (2024). The use of Argo floats as virtual moorings for monitoring the South Baltic Sea. *Oceanologia* **66** 99–110.

OLSON, S., JANSEN, M. F., ABBOT, D. S., HALEVY, I. and GOLDBLATT, C. (2022). The effect of ocean salinity on climate and its implications for Earth's habitability. *Geophys. Res. Lett.* **49** e2021GL095748.

PAWLOWICZ, R. (2013). Key physical variables in the ocean: Temperature, salinity, and density. *Nat. Educ. Knowl.* **4**.

ROEMMICH, D. and GILSON, J. (2009). The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo program. *Prog. Oceanogr.* **82** 81–100.

SALVAÑA, M. L, CAO, J. and JUN, M. (2026). Supplement to "3D bivariate spatial modelling of Argo ocean temperature and salinity profiles." https://doi.org/10.1214/26-AOAS2149SUPP

SAMBE, F. and SUGA, T. (2022). Unsupervised clustering of Argo temperature and salinity profiles in the mid-latitude northwest Pacific Ocean and revealed influence of the Kuroshio extension variability on the vertical structure distribution. *J. Geophys. Res., Oceans* **127** e2021JC018138.

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

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

WALKER, R. H., SMITH, G. D., HUDSON, S. B., FRENCH, S. S. and WALTERS, A. W. (2020). Warmer temperatures interact with salinity to weaken physiological facilitation to stress in freshwater fishes. *Conserv. Physiol.* **8** coaa107.

WANG, T., GILLE, S. T., MAZLOFF, M. R., ZILBERMAN, N. V. and DU, Y. (2020). Eddy-induced acceleration of Argo floats. *J. Geophys. Res., Oceans* **125** e2019JC016042.

WONG, A. P., WIJFFELS, S. E., RISER, S. C., POULIQUEN, S., HOSODA, S., ROEMMICH, D., GILSON, J., JOHNSON, G. C., MARTINI, K. et al. (2020). Argo data 1999–2019: Two million temperature-salinity profiles and subsurface velocity observations from a global array of profiling floats. *Front. Mar. Sci.* **700**.

XIE, S.-P., KUNITANI, T., KUBOKAWA, A., NONAKA, M. and HOSODA, S. (2000). Interdecadal thermocline variability in the North Pacific for 1958–97: A GCM simulation. *J. Phys. Oceanogr.* **30** 2798–2813.

YARGER, D., STOEV, S. and HSING, T. (2022). A functional-data approach to the Argo data. *Ann. Appl. Stat.* **16** 216–246. MR4400508 https://doi.org/10.1214/21-aoas1477

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1148–1167
https://doi.org/10.1214/26-AOAS2177

A TIME WARPING MODEL FOR SEASONAL DATA WITH APPLICATION TO AGE ESTIMATION FROM NARWHAL TUSKS

BY LARS N. REITER[1,a], ADAM G. HOFFMANN[1,b], MADS PETER HEIDE-JØRGENSEN[2,d], EVA GARDE[2,e], ADELINE SAMSON[3,f] AND SUSANNE DITLEVSEN[1,c]

[1]*Department of Mathematical Sciences, University of Copenhagen,* [a]*lrn@math.ku.dk,* [b]*adam.hoffmann@sund.ku.dk,* [c]*susanne@math.ku.dk*

[2]*Greenland Institute of Natural Resources,* [d]*mhj@mail.ghsdk.dk,* [e]*evga@mail.ghsdk.dk*

[3]*Jean Kuntzmann Laboratory, Université Grenoble Alpes,* [f]*adeline.leclercq-samson@univ-grenoble-alpes.fr*

Signals with varying periodicity frequently appear in real-world phenomena, necessitating the development of efficient modelling techniques to map the measured nonlinear timeline to linear time. Here we propose a regression model that allows for a representation of periodic and dynamic patterns observed in time series data. The model incorporates a hidden strictly positive stochastic process that represents the instantaneous frequency, allowing the model to adapt and accurately capture varying time scales. A case study focusing on age estimation of narwhal tusks is presented, where cyclic element signals associated with annual growth layer groups are analyzed. We apply the methodology to data from one such tusk collected in West Greenland and use the fitted model to estimate the age of the narwhal. The proposed method is validated using simulated signals with known cycle counts and practical considerations and modelling challenges are discussed in detail. This research contributes to the field of time series analysis, providing a tool and valuable insights for understanding and modeling complex cyclic patterns in diverse domains.

REFERENCES

AMAIS, R. S., MOREAU, P. S., FRANCISCHINI, D. S., MAGNUSSON, R., LOCOSSELLI, G. M., GODOY-VEIGA, M., CECCANTINI, G., RODRIGUEZ, D. R. O., TOMAZELLO-FILHO, M. et al. (2021). Trace elements distribution in tropical tree rings through high-resolution imaging using LA-ICP-MS analysis. *J. Trace Elem. Med. Biol.* **68** 126872.

BIBBY, B. M. and SØRENSEN, M. (1995). Martingale estimation functions for discretely observed diffusion processes. *Bernoulli* **1** 17–39. MR1354454 https://doi.org/10.2307/3318679

CHARLES, C. D., LYNCH-STIEGLITZ, J., NINNEMANN, U. S. and FAIRBANKS, R. G. (1996). Climate connections between the hemisphere revealed by deep sea sediment core/ice core correlations. *Earth Planet. Sci. Lett.* **142** 19–27.

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

COX, J. C., INGERSOLL, J. E. JR. and ROSS, S. A. (1985). A theory of the term structure of interest rates. *Econometrica* **53** 385–407. MR0785475 https://doi.org/10.2307/1911242

DAGUM, E. (2013). Time series modelling and decomposition. *Statistica* **70**. https://doi.org/10.6092/issn.1973-2201/3597

DAHLHAUS, R., DUMONT, T., LE CORFF, S. and NEDDERMEYER, J. C. (2017). Statistical inference for oscillation processes. *Statistics* **51** 61–83. MR3600462 https://doi.org/10.1080/02331888.2016.1266985

DEL MORAL, P., DOUCET, A. and JASRA, A. (2006). Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 411–436. MR2278333 https://doi.org/10.1111/j.1467-9868.2006.00553.x

DELYON, B., LAVIELLE, M. and MOULINES, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.* **27** 94–128. MR1701103 https://doi.org/10.1214/aos/1018031103

Key words and phrases. Sclerochronology, trace element analysis, stochastic phase process, warping of signals from distance to time, age estimation, detecting seasonality, narwhal tusk.

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, Methodol.* **39** 1–38. MR0501537

DIETZ, R., DESFORGES, J.-P., RIGÉT, F. F., AUBAIL, A., GARDE, E., AMBUS, P., DRIMMIE, R., HEIDE-JØRGENSEN, M. P. and SONNE, C. (2021). Analysis of narwhal tusks reveals lifelong feeding ecology and Mercury exposure. *Curr. Biol.* **31** 2012–2019.e2. https://doi.org/10.1016/j.cub.2021.02.018

DITLEVSEN, S., RUBIO, A. C. and LANSKY, P. (2020). Transient dynamics of Pearson diffusions facilitates estimation of rate parameters. *Commun. Nonlinear Sci. Numer. Simul.* **82** 105034, 15. MR4019857 https://doi.org/10.1016/j.cnsns.2019.105034

DITLEVSEN, S. and LANSKY, P. (2006). Estimation of the input parameters in the Feller neuronal model. *Phys. Rev. E* **73** 061910, 9. MR2276280 https://doi.org/10.1103/PhysRevE.73.061910

DITLEVSEN, S. and SAMSON, A. (2014). Estimation in the partially observed stochastic Morris-Lecar neuronal model with particle filter and stochastic approximation methods. *Ann. Appl. Stat.* **8** 674–702. MR3262530 https://doi.org/10.1214/14-AOAS729

DITLEVSEN, S. and SØRENSEN, M. (2004). Inference for observations of integrated diffusion processes. *Scand. J. Stat.* **31** 417–429. MR2087834 https://doi.org/10.1111/j.1467-9469.2004.02_023.x

DOUCET, A., DE FREITAS, N. and GORDON, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo Methods in Practice*. *Stat. Eng. Inf. Sci.* 3–14. Springer, New York. MR1847784 https://doi.org/10.1007/978-1-4757-3437-9_1

FORMAN, J. L. and SØRENSEN, M. (2008). The Pearson diffusions: A class of statistically tractable diffusion processes. *Scand. J. Stat.* **35** 438–465. MR2446729 https://doi.org/10.1111/j.1467-9469.2007.00592.x

GARDE, E., DITLEVSEN, S., OLSEN, J. and HEIDE-JØRGENSEN, M. P. (2024). A radiocarbon bomb pulse model for estimating the age of North Atlantic cetaceans. *Biol. Lett.* **20** 20240350. https://doi.org/10.1098/rsbl.2024.0350

GARDE, E., HEIDE-JØRGENSEN, M. P., HANSEN, S. H., NACHMAN, G. and FORCHHAMMER, M. C. (2007). Age-specific growth and remarkable longevity in narwhals (Monodon monoceros) from West Greenland as estimated by aspartic acid racemization. *J. Mammal.* **88** 49–58.

GARDE, E., TERVO, O. M., SINDING, M.-H. S., NIELSEN, N. H., CORNETT, C. and HEIDE-JØRGENSEN, M. P. (2022). Biological parameters in a declining population of narwhals (Monodon monoceros) in Scoresby Sound, Southeast Greenland. *Antarct. Sci.* **8** 329–348.

GLASS, L. and MACKEY, M. C. (1988). *From Clocks to Chaos*: *The Rhythms of Life*. Princeton Univ. Press, Princeton, NJ. MR0952149

HAY, K. A. (1980). Age determination of the narwhal, Monodon monoceros L. *Rep. Int. Whal. Comm.* 119–132.

HEIDE-JØRGENSEN, M. P., BLACKWELL, S. B., WILLIAMS, T. M., SINDING, M. H. S., SKOVRIND, M., TERVO, O. M., GARDE, E., HANSEN, R. G., NIELSEN, N. H. et al. (2020). Some like it cold: Temperature-dependent habitat selection by narwhals. *Ecol. Evol.* **10** 8073–8090. https://doi.org/10.1002/ece3.6464

HEIMBRAND, Y., LIMBURG, K. E., HÜSSY, K., CASINI, M., SJÖBERG, R., PALMÉN BRATT, A.-M., LEVINSKY, S.-E., KARPUSHEVSKAIA, A., RADTKE, K. et al. (2020). Seeking the true time: Exploring otolith chemistry as an age-determination tool. *J. Fish Biol.* **97** 552–565. https://doi.org/10.1111/jfb.14422

HÜSSY, K., KRÜGER-JOHNSEN, M., THOMSEN, T. B., HEREDIA, B. D., NÆRAA, T., LIMBURG, K. E., HEIMBRAND, Y., MCQUEEN, K., HAASE, S. et al. (2021a). It's elemental, my dear Watson: Validating seasonal patterns in otolith chemical chronologies. *Can. J. Fish. Aquat. Sci.* **78** 551–566.

HÜSSY, K., LIMBURG, K. E., DE PONTUAL, H., THOMAS, O. R., COOK, P. K., HEIMBRAND, Y., BLASS, M. and STURROCK, A. M. (2021b). Trace element patterns in otoliths: The role of biomineralization. *Reviews Fish. Sci. Aquac.* **29** 445–477.

HÜSSY, K. and MOSEGAARD, H. (2004). Atlantic cod (Gadus morhua) growth and otolith accretion characteristics modelled in a bioenergetics context. *Can. J. Fish. Aquat. Sci.* **61** 1021–1031.

JANK, W. (2006). Implementing and diagnosing the stochastic approximation EM algorithm. *J. Comput. Graph. Statist.* **15** 803–829. MR2297632 https://doi.org/10.1198/106186006X157469

JEONG, Y.-S., JEONG, M. K. and OMITAOMU, O. A. (2011). Weighted dynamic time warping for time series classification. *Pattern Recognit.* **44** 2231–2240.

JOUZEL, J. and MASSON-DELMOTTE, V. (2010). Paleoclimates: What do we learn from deep ice cores? *Wiley Interdiscip. Rev.*: *Clim. Change* **1** 654–669.

KING, A. A., NGUYEN, D. and IONIDES, E. L. (2016). Statistical inference for partially observed Markov processes via the R package pomp. *J. Stat. Softw.* **69** 1–43.

KOCH, J. and GÜNTHER, D. (2011). Review of the state-of-the-art of laser ablation inductively coupled plasma mass spectrometry. *Appl. Spectrosc.* **65** 155A–162A.

LAIDRE, K. L. and HEIDE-JØRGENSEN, M. P. (2005). Winter feeding intensity of narwhals (Monodon Monoceros). *Mar. Mamm. Sci.* **21** 45–57. https://doi.org/10.1111/j.1748-7692.2005.tb01207.x

LAVIELLE, M. (2015). *Mixed Effects Models for the Population Approach*: *Models*, *Tasks*, *Methods and Tools*. *Chapman & Hall/CRC Biostatistics Series*. CRC Press, Boca Raton, FL. MR3331127

LIMBURG, K., OLSON, C., WALTHER, Y., DALE, D., SLOMP, C. P. and HØIE, H. (2011). Tracking Baltic hypoxia and cod migration over millennia with natural tags. *Proc. Natl. Acad. Sci. USA* **108** E177–82. https://doi.org/10.1073/pnas.1100684108

LIXOFT SAS (2024). SAEM Lixoft SAS Monolix Suite 2024R1. Available at https://monolixsuite.slp-software.com/monolix/2024R1/saem (accessed April 16, 2025).

MONSERUD, R. A. and MARSHALL, J. D. (2001). Time-series analysis of δ13C from tree rings. I. Time trends and autocorrelation. *Tree Physiol.* **21** 1087–1102.

MOSBACHER, J. B., MICHELSEN, A., STELVIG, M., HENDRICHSEN, D. K. and SCHMIDT, N. M. (2016). Show me your rump hair and I will tell you what you ate–the dietary history of muskoxen (Ovibos moschatus) revealed by sequential stable isotope analysis of guard hairs. *PLoS ONE* **11** e0152874.

MOSEGAARD, H., SVEDÄNG, H. and TABERMAN, K. (1988). Uncoupling of somatic and otolith growth rates in Arctic Char (Salvelinus alpinus) as an effect of differences in temperature response. *Can. J. Fish. Aquat. Sci.* **45** 1514–1524. https://doi.org/10.1139/f88-180

NAPOLITANO, A. and GARDNER, W. A. (2016). Algorithms for analysis of signals with time-warped cyclostationarity. In 2016 50*th Asilomar Conference on Signals*, *Systems and Computers* 539–543. IEEE.

OSGOOD, B. (2007). The Fourier Transform and its applications. Stanford University.

READ, F. L., HOHN, A. A. and LOCKYER, C. H. (2018). A review of age estimation methods in marine mammals with special reference to monodontids.

REITER, L. N, HOFFMANN, A. G, HEIDE-JØRGENSEN, M. P, GARDE, E., SAMSON, A. and DITLEVSEN, S. (2026). Supplement to "A time warping model for seasonal data with application to age estimation from narwhal tusks." https://doi.org/10.1214/26-AOAS2177SUPPA, https://doi.org/10.1214/26-AOAS2177SUPPB

REITER, L. N., THOMSEN, T. B., HEREDIA, B. D., SAMSON, A., WIELANDT, D. K. P., HEIDE-JØRGENSEN, M. P., DITLEVSEN, S. and GARDE, E. (2025). Chronicles in ivory: Estimating the age of narwhals (Monodon monoceros) through stochastic modeling of seasonally varying trace elements. *Front. Mar. Sci.* **12**.

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1190470 https://doi.org/10.1002/9780470316856

SHUMWAY, R. H. and STOFFER, D. S. (2017). *Time Series Analysis and Its Applications*: *With R Examples*, 4th ed. *Springer Texts in Statistics*. Springer, Cham. MR3642322 https://doi.org/10.1007/978-3-319-52452-8

SØRENSEN, M. (2012). Estimating functions for diffusion-type processes. In *Statistical Methods for Stochastic Differential Equations*. *Monogr. Statist. Appl. Probab.* **124** 1–107. CRC Press, Boca Raton, FL. MR2976982 https://doi.org/10.1201/b12126-2

STOUNBERG, J., THOMSEN, T. B., HEREDIA, B. D. and HÜSSY, K. (2022). Eyes and ears: A comparative approach linking the chemical composition of cod otoliths and eye lenses. *J. Fish Biol.* **101** 985–995.

THOMSEN, T., KEULEN, N., HEREDIA, B. and MALKKI, S. (2022). Cannon balls from Bremervörde (D). A characterization of historic cannon balls using SEM and LA-ICPMS microchemistry. Danmarks og Grønlands Geologiske Undersøgelse Rapport 2022. GEUS. https://doi.org/10.22008/gpub/34640

WANG, Y., MILLER, D. J., POSKANZER, K., WANG, Y., TIAN, L. and YU, G. (2016). Graphical time warping for joint alignment of multiple curves. *Adv. Neural Inf. Process. Syst.* **29**.

WATT, C. and FERGUSON, S. (2014). Fatty acids and stable isotopes (δ13C and δ15N) reveal temporal changes in narwhal (Monodon monoceros) diet linked to migration patterns. *Mar. Mamm. Sci.* **31**. https://doi.org/10.1111/mms.12131

WATT, C. A., STEWART, B. E., LOSETO, L., HALLDORSON, T. and FERGUSON, S. H. (2020). Estimating narwhal (Monodon monoceros) age using tooth layers and aspartic acid racemization of eye lens nuclei. *Mar. Mamm. Sci.* **36** 103–115.

ZHAO, S.-T., MATTHEWS, C. J. and WATT, C. A. (2025). δ15N and δ13C cycles in narwhal (Monodon monoceros) embedded teeth reveal seasonal variation in ecology and/or physiology. *R. Soc. Open Sci.* **12** 242237.

ZOPPI, U., SKOPEC, Z., SKOPEC, J., JONES, G., FINK, D., HUA, Q., JACOBSEN, G., TUNIZ, C. and WILLIAMS, A. (2004). Forensic applications of 14C bomb-pulse dating. *Nucl. Instrum. Methods Phys. Res., Sect. B, Beam Interact. Mater. Atoms* **223** 770–775.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1168–1186
https://doi.org/10.1214/26-AOAS2159

SPATIAL PREDICTION OF LOCAL SOIL EROSION DISTRIBUTION IN THE WASSERSTEIN SPACE

BY JIAMING QIU[1,a], XIONGTAO DAI[2,b], ZHENGYUAN ZHU[3,c] AND SHUIQING YIN[4,d]

[1]*Public Health Sciences Division, Fred Hutchinson Cancer Center,* [a]*jqiu3@fredhutch.org*
[2]*Division of Biostatistics, University of California, Berkeley,* [b]*xdai@berkeley.edu*
[3]*Department of Statistics, Iowa State University,* [c]*zhuz@iastate.edu*
[4]*State Key Laboratory of Earth Surface Processes and Hazards Risk Governance (ESPHR), Faculty of Geographical Science, Beijing Normal University,* [d]*yinshuiqing@bnu.edu.cn*

Obtaining precise erosion measurements requires costly fieldwork, making it infeasible to directly survey large domains such as a province or river basin. To extend fieldwork results across such extensive domains, we propose a novel spatial prediction method that treats local erosion distributions as objects in the Wasserstein space. These distributions are mapped into square-integrable trajectories and represented via basis expansion, forming a multivariate random field that captures spatial dependence. By applying local regression and Kriging in this representation, our approach flexibly models and predicts erosion distributions at arbitrary locations. This framework improves prediction for functionals of the distribution, such as the mean and exceedance probabilities. Simulation studies demonstrate that the proposed method outperforms a misspecified parametric alternative and existing Fréchet regression approaches. We illustrate the approach with a detailed erosion analysis in Shaanxi Province, China, where local measurements from surveyed watersheds are extended to predict erosion distributions across the entire province using covariates such as land use and elevation.

REFERENCES

BALZANELLA, A. and IRPINO, A. (2020). Spatial prediction and spatial dependence monitoring on georeferenced data streams. *Stat. Methods Appl.* **29** 101–128. MR4076121 https://doi.org/10.1007/s10260-019-00462-0

BHATTACHARJEE, S. and MÜLLER, H.-G. (2023). Single index Fréchet regression. *Ann. Statist.* **51** 1770–1798. MR4658576 https://doi.org/10.1214/23-aos2307

BIGOT, J., GOUET, R., KLEIN, T. and LÓPEZ, A. (2017). Geodesic PCA in the Wasserstein space by convex PCA. *Ann. Inst. Henri Poincaré Probab. Stat.* **53** 1–26. MR3606732 https://doi.org/10.1214/15-AIHP706

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ, USA. https://doi.org/10.1002/9781119115151

DUNSON, D. B. and PARK, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* **95** 307–323. MR2521586 https://doi.org/10.1093/biomet/asn012

FARR, T. G. and KOBRICK, M. (2000). Shuttle radar topography mission produces a wealth of data. *Eos Trans. AGU* **81** 583–585. https://doi.org/10.1029/EO081i048p00583

GELFAND, A. E., ed. (2010). *Handbook of Spatial Statistics*. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL.

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

GHOSAL, A., MEIRING, W. and PETERSEN, A. (2023). Fréchet single index models for object response regression. *Electron. J. Stat.* **17** 1074–1112. MR4575027 https://doi.org/10.1214/23-ejs2120

HE, X. (1997). Quantile curves without crossing. *Amer. Statist.* **51** 186–192. https://doi.org/10.1080/00031305.1997.10473959

HRON, K., MENAFOGLIO, A., TEMPL, M., HRŮZOVÁ, K. and FILZMOSER, P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Comput. Statist. Data Anal.* **94** 330–350. MR3412829 https://doi.org/10.1016/j.csda.2015.07.007

Key words and phrases. Wasserstein space, functional data analysis, Kriging, zero-inflated model.

KOENKER, R., CHERNOZHUKOV, V., HE, X. and PENG, L., eds. (2017). *Handbook of Quantile Regression*. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR3728340

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

LIU, B., GUO, S., LI, Z., XIE, Y., ZHANG, K. and LIU, X. (2013). Sample survey on water erosion in China. *Soil Water Conserv. China* **10** 26–34.

LIU, B. Y., ZHANG, K. L. and XIE, Y. (2002). An empirical soil loss equation. In *Process of Soil Erosion and Its Environment Effect II* 21–25. Tsinghua Univ. Press, Beijing.

LIU, J., KUANG, W., ZHANG, Z., XU, X., QIN, Y., NING, J., ZHOU, W., ZHANG, S., LI, R. et al. (2014). Spatiotemporal characteristics, patterns, and causes of land-use changes in China since the late 1980s. *J. Geogr. Sci.* **24** 195–210. https://doi.org/10.1007/s11442-014-1082-6

MATEU, J. and GIRALDO, R., eds. (2021). *Geostatistical Functional Data Analysis*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ.

MENAFOGLIO, A. (2021). Spatial statistics for distributional data in Bayes spaces: From object-oriented Kriging to the analysis of warping functions. In *Advances in Compositional Data Analysis—Festschrift in Honour of Vera Pawlowsky-Glahn* (P. Filzmoser, K. Hron, J. A. Martín-Fernández and J. Palarea-Albaladejo, eds.) 207–224. Springer, Cham. MR4299317 https://doi.org/10.1007/978-3-030-71175-7_11

MENAFOGLIO, A., GUADAGNINI, A. and SECCHI, P. (2014). A Kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stoch. Environ. Res. Risk Assess.* **28** 1835–1851. https://doi.org/10.1007/s00477-014-0849-8

PANARETOS, V. M. and ZEMEL, Y. (2020). *An Invitation to Statistics in Wasserstein Space*. *SpringerBriefs in Probability and Mathematical Statistics*. Springer, Cham. MR4350694 https://doi.org/10.1007/978-3-030-38438-8

PETERSEN, A. and MÜLLER, H.-G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *Ann. Statist.* **44** 183–218. MR3449766 https://doi.org/10.1214/15-AOS1363

PETERSEN, A. and MÜLLER, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *Ann. Statist.* **47** 691–719. MR3909947 https://doi.org/10.1214/17-AOS1624

PETERSEN, A., ZHANG, C. and KOKOSZKA, P. (2022). Modeling probability density functions as data objects. *Econom. Stat.* **21** 159–178. MR4366852 https://doi.org/10.1016/j.ecosta.2021.04.004

PIGOLI, D., MENAFOGLIO, A. and SECCHI, P. (2016). Kriging prediction for manifold-valued random fields. *J. Multivariate Anal.* **145** 117–131. MR3459942 https://doi.org/10.1016/j.jmva.2015.12.006

QIU, J., DAI, X., ZHU, Z. and YIN, S. (2026). Supplement to "Spatial prediction of local soil erosion distribution in the Wasserstein space." https://doi.org/10.1214/26-AOAS2159SUPPA, https://doi.org/10.1214/26-AOAS2159SUPPB

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

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

TUCKER, D. C. and WU, Y. (2025). Partially-global Fréchet regression. *Statist. Sinica* **35** 713–736. MR4882560

VAN DEN BOOGAART, K. G., EGOZCUE, J. J. and PAWLOWSKY-GLAHN, V. (2014). Bayes Hilbert spaces. *Aust. N. Z. J. Stat.* **56** 171–194. MR3226435 https://doi.org/10.1111/anzs.12074

XU, S. G. and REICH, B. J. (2023). Bayesian nonparametric quantile process regression and estimation of marginal quantile effects. *Biometrics* **79** 151–164. MR4572512 https://doi.org/10.1111/biom.13576

YIN, S.-Q., WANG, Z., ZHU, Z., ZOU, X.-K. and WANG, W.-T. (2018). Using Kriging with a heterogeneous measurement error to improve the accuracy of extreme precipitation return level estimation. *J. Hydrol.* **562** 518–529. https://doi.org/10.1016/j.jhydrol.2018.04.064

ZHANG, H. and LI, Y. (2022). Unified principal component analysis for sparse and dense functional data under spatial dependency. *J. Bus. Econom. Statist.* **40** 1523–1537. MR4492051 https://doi.org/10.1080/07350015.2021.1938085

ZHEN, L. (2013). The national census for soil erosion and dynamic analysis in China. *Int. Soil Water Conserv. Res.* **1** 12–18. https://doi.org/10.1016/S2095-6339(15)30035-6

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1187–1207
https://doi.org/10.1214/26-AOAS2194

INCORPORATING CORRELATED NUGGET EFFECTS IN MULTIVARIATE SPATIAL MODELS: AN APPLICATION TO ARGO OCEAN DATA

BY DAMILYA SADUAKHAS[1,a], DAVID BOLIN[1,b], XIAOTIAN JIN[1,c], ALEXANDRE B. SIMAS[1,d] AND JONAS WALLIN[2,e]

[1]*Statistics Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST),* [a]*damilya.saduakhas@kaust.edu.sa,* [b]*david.bolin@kaust.edu.sa,* [c]*xiaotian.jin@kaust.edu.sa,* [d]*alexandre.simas@kaust.edu.sa*

[2]*Department of Statistics, Lund University,* [e]*jonas.wallin@stat.lu.se*

Accurate analysis of global oceanographic data, such as temperature and salinity profiles from the Argo program, requires geostatistical models that capture complex spatial dependencies. We propose Gaussian and non-Gaussian hierarchical multivariate Matérn-SPDE models with correlated nugget effects that jointly account for small-scale variability and measurement-error correlations between co-located observations. Simulations show that ignoring such correlations biases cross-variable dependence estimates, while incorporating them improves parameter recovery. Applied to 14 years of global Argo data, our models yield lower temperature–salinity dependence than models that assume independent noise, indicating that standard approaches tend to overstate this dependence. Cross-validation shows that the Gaussian model with correlated noise achieves the best point predictions, while the non-Gaussian (NIG) model with independent noise yields better-calibrated probabilistic scores. These findings highlight the importance of relaxing the independent-noise assumption in multivariate hierarchical models.

REFERENCES

ARGO (2023). Argo float data and metadata from global data assembly centre (Argo GDAC)—Snapshot of Argo GDAC of August 2023. SEANOE. https://doi.org/10.17882/42182#104145

ASAR, Ö., BOLIN, D., DIGGLE, P. J. and WALLIN, J. (2020). Linear mixed effects models for non-Gaussian continuous repeated measurement data. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **69** 1015–1065. MR4166856 https://doi.org/10.1111/rssc.12405

BARKER, P. M., DUNN, J. R., DOMINGUES, C. M. and WIJFFELS, S. E. (2011). Pressure sensor drifts in Argo and their impacts. *J. Atmos. Ocean. Technol.* **28** 1036–1049. https://doi.org/10.1175/2011JTECHO831.1

BARNDORFF-NIELSEN, O. E. (1997). Normal inverse Gaussian distributions and stochastic volatility modelling. *Scand. J. Stat.* **24** 1–13. MR1436619 https://doi.org/10.1111/1467-9469.00045

BOLIN, D. (2014). Spatial Matérn fields driven by non-Gaussian noise. *Scand. J. Stat.* **41** 557–579. MR3249417 https://doi.org/10.1111/sjos.12046

BOLIN, D., JIN, X., SIMAS, A. and WALLIN, J. (2025). ngme2: Latent Mixed Effects Models with Flexible Distributions R package version 0.7.1. https://github.com/davidbolin/ngme2.

BOLIN, D. and KIRCHNER, K. (2020). The rational SPDE approach for Gaussian random fields with general smoothness. *J. Comput. Graph. Statist.* **29** 274–285. MR4116041 https://doi.org/10.1080/10618600.2019.1665537

BOLIN, D., SIMAS, A. B. and XIONG, Z. (2024). Covariance-based rational approximations of fractional SPDEs for computationally efficient Bayesian inference. *J. Comput. Graph. Statist.* **33** 64–74. MR4713943 https://doi.org/10.1080/10618600.2023.2231051

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

BOLIN, D. and WALLIN, J. (2023). Local scale invariance and robustness of proper scoring rules. *Statist. Sci.* **38** 140–159. MR4534647 https://doi.org/10.1214/22-sts864

Key words and phrases. Non-Gaussian random fields, SPDE approach, Argo project, multivariate random fields, nugget effect.

BRETHERTON, F. P., DAVIS, R. E. and FANDRY, C. B. (1976). A technique for objective analysis and design of oceanographic experiments applied to MODE-73. *Deep-Sea Res. Oceanogr. Abstr.* **23** 559–582. https://doi.org/10.1016/0011-7471(76)90001-2

CABANES, C., THIERRY, V. and LAGADEC, C. (2016). Improvement of bias detection in Argo float conductivity sensors and its application in the North Atlantic. Deep Sea research Part I. *Deep-Sea Res., Part* 1, *Oceanogr. Res. Pap.* **114** 128–136. https://doi.org/10.1016/j.dsr.2016.05.007

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

DURACK, P. J., WIJFFELS, S. E. and MATEAR, R. J. (2012). Ocean salinities reveal strong global water cycle intensification during 1950 to 2000. *Science* **336** 455–458. https://doi.org/10.1126/science.1212222

FUGLSTAD, G.-A., SIMPSON, D., LINDGREN, F. and RUE, H. (2015). Does non-stationary spatial data always require non-stationary random fields? *Spat. Stat.* **14** 505–531. MR3431054 https://doi.org/10.1016/j.spasta.2015.10.001

GAILLARD, F., AUTRET, E., THIERRY, V., GALAUP, P., COATANOAN, C. and LOUBRIEU, T. (2009). Quality control of large Argo datasets. *J. Atmos. Ocean. Technol.* **26** 337–351. https://doi.org/10.1175/2008JTECHO552.1

GAILLARD, F., REYNAUD, T., THIERRY, V., KOLODZIEJCZYK, N. and VON SCHUCKMANN, K. (2016). In-situ based reanalysis of the global ocean temperature and salinity with ISAS: Variability of the heat content and steric height. *J. Climate* **29** 1305–1323. https://doi.org/10.1175/JCLI-D-15-0028.1

GIGLIO, D., LYUBCHICH, V. and MAZLOFF, M. R. (2018). Estimating oxygen in the Southern Ocean using Argo temperature and salinity. *J. Geophys. Res., Oceans* **123** 4280–4297. https://doi.org/10.1029/2017jc013404

GNEITING, T., KLEIBER, W. and SCHLATHER, M. (2010). Matérn cross-covariance functions for multivariate random fields. *J. Amer. Statist. Assoc.* **105** 1167–1177. MR2752612 https://doi.org/10.1198/jasa.2010.tm09420

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

HALL, A. and MANABE, S. (1997). Can local linear stochastic theory explain sea surface temperature and salinity variability? *Clim. Dyn.* **13** 167–180.

HOSODA, S., OHIRA, T. and NAKAMURA, T. (2008). A monthly mean dataset of global Oceanic temperature and salinity derived from Argo float observations. *JAMSTEC Rep. Res. Dev.* **8** 47–59. https://doi.org/10.5918/jamstecr.8.47

HU, X., SIMPSON, D., LINDGREN, F. and RUE, H. (2013). Multivariate Gaussian random fields using systems of stochastic partial differential equations.

HU, X. and STEINSLAND, I. (2016). Spatial modeling with system of stochastic partial differential equations. *Wiley Interdiscip. Rev.: Comput. Stat.* **8** 112–125. MR3466001 https://doi.org/10.1002/wics.1378

JAYNE, S., ROEMMICH, D., ZILBERMAN, N., RISER, S., JOHNSON, K., JOHNSON, G. and PIOTROWICZ, S. (2017). The Argo program: Present and future. *Oceanography* **30** 18–28. https://doi.org/10.5670/oceanog.2017.213

KELLEY, D., RICHARDS, C. and SCOR/IAPSO, W. (2024). gsw: Gibbs Sea Water Functions R package version 1.2-0. https://doi.org/10.32614/CRAN.package.gsw

KIDO, S., NONAKA, M. and TANIMOTO, Y. (2021). Sea surface temperature–salinity covariability and its scale-dependent characteristics. *Geophys. Res. Lett.* **48** e2021GL096010. https://doi.org/10.1029/2021GL096010

KUUSELA, M. and STEIN, M. L. (2018). Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proc. R. Soc. A, Math. Phys. Eng. Sci.* **474** 20180400. https://doi.org/10.1098/rspa.2018.0400

LINDGREN, F. (2025). fmesher: Triangle Meshes and Related Geometry Tools R package version 0.3.0.9012.

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

LIU, C., LIANG, X., PONTE, R. M. and CHAMBERS, D. P. (2024). "Salty drift" of Argo floats affects the gridded ocean salinity products. *J. Geophys. Res., Oceans* **129** e2023JC020871. https://doi.org/10.1029/2023JC020871

MATÉRN, B. (1960). *Spatial Variation*: *Stochastic Models and Their Application to Some Problems in Forest Surveys and Other Sampling Investigations*. Statens Skogsforskningsinstitut, Stockholm. Meddelanden Fran Statens Skogsforskningsinstitut, Band 49, Nr. 5. MR0169346

MATHERON, G. (1979). Recherche de simplification dans un problème de cokrigeage. Centre de Géostatistique, Ecole des Mines de Paris, Fontainebleau. Publication N-628.

PARK, B., KUUSELA, M., GIGLIO, D. and GRAY, A. (2023). Spatiotemporal local interpolation of global ocean heat transport using Argo floats: A debiased latent Gaussian process approach. *Ann. Appl. Stat.* **17** 1491–1520. MR4582722 https://doi.org/10.1214/22-aoas1679

PODGÓRSKI, K. and WALLIN, J. (2016). Convolution-invariant subclasses of generalized hyperbolic distributions. *Comm. Statist. Theory Methods* **45** 98–103. MR3440372 https://doi.org/10.1080/03610926.2013.821489

RISER, S. C., FREELAND, H. J., ROEMMICH, D., WIJFFELS, S., TROISI, A., BELBÉOCH, M., GILBERT, D., XU, J., POULIQUEN, S. et al. (2016). Fifteen years of ocean observations with the global Argo array. *Nat. Clim. Change* **6** 145–153.

ROEMMICH, D. and GILSON, J. (2009). The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo program. *Prog. Oceanogr.* **82** 81–100. https://doi.org/10.1016/j.pocean.2009.03.004

SADUAKHAS, D., BOLIN, D., JIN, X., SIMAS, A. B and WALLIN, J. (2026). Supplement to "Incorporating correlated nugget effects in multivariate spatial models: An application to Argo ocean data." https://doi.org/10.1214/26-AOAS2194SUPPA, https://doi.org/10.1214/26-AOAS2194SUPPB

SALVANA, M. L., CAO, J. and JUN, M. (2025). 3*D Bivariate Spatial Modelling of Argo Ocean Temperature and Salinity Profiles*.

TALLEY, L. D., PICKARD, G. L., EMERY, W. J. and SWIFT, J. H. (2011). Chapter 3—physical properties of seawater. In *Descriptive Physical Oceanography* 29–65 6th ed. Academic Press, Boston, MA. https://doi.org/10.1016/B978-0-7506-4552-2.10003-4

THE MATHWORKS, I. (2022). Natick, Massachusetts, United States. Matlab version: 9.12.0 (R2022a).

WACKERNAGEL, H. (2003). *Multivariate Geostatistics*: *An Introduction with Applications*, 3rd ed. ed. Springer, Berlin. https://doi.org/10.1007/978-3-662-05294-5

WALLIN, J. and BOLIN, D. (2015). Geostatistical modelling using non-Gaussian Matérn fields. *Scand. J. Stat.* **42** 872–890. MR3391697 https://doi.org/10.1111/sjos.12141

WHITTLE, P. (1963). Stochastic processes in several dimensions. *Bull. Inst. Int. Stat.* **40** 974–994. MR0173287

WONG, A. P. S., GILSON, J. and CABANES, C. (2023). Argo salinity: Bias and uncertainty evaluation. *Earth Syst. Sci. Data* **15** 383–393. https://doi.org/10.5194/essd-15-383-2023

WONG, A. P. S., WIJFFELS, S. E., RISER, S. C., POULIQUEN, S., HOSODA, S., ROEMMICH, D. et al. (2020). Argo data 1999–2019: Two million temperature-salinity profiles and subsurface velocity observations from a global array of profiling floats. *Front. Mar. Sci.* **7**. https://doi.org/10.3389/fmars.2020.00700

YARGER, D., STOEV, S. and HSING, T. (2022). A functional-data approach to the Argo data. *Ann. Appl. Stat.* **16** 216–246. MR4400508 https://doi.org/10.1214/21-aoas1477

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1208–1226
https://doi.org/10.1214/26-AOAS2188

LÉVY PROCESSES FOR JUMPING GROWTH OF SPINY LOBSTERS, *Panulirus ornatus*

BY YOU-GAN WANG[1,a] AND CHUAN HUI FOO[2,b]

[1]*School of Statistics and Data Science, Guangdong University of Finance and Economics,* [a]*wangyg@gdufe.edu.cn*
[2]*Department of Mathematics, Sultan Idris Education University,* [b]*foo.ch@fsmt.upsi.edu.my*

The discontinuous moulting process in crustaceans poses fundamental challenges for growth modelling and can lead to biologically implausible estimates of asymptotic size under traditional continuous-growth frameworks such as the von Bertalanffy curve. We develop a stochastic growth model that jointly characterises moult increment (MI) and intermoult period (IP) through a unified convolution-based likelihood. Individual growth is represented within a Lévy-inspired jump framework that enforces monotone but discontinuous size trajectories by modelling MI through a beta-type jump process with biologically realistic support, while IP is described by a gamma generalised linear model. This construction yields a joint likelihood for MI and IP and permits maximum likelihood estimation together with profile-likelihood inference for parameters governing both the size and timing of moults.

In an application to tank data on *Panulirus ornatus*, we compare several plausible jump and waiting-time distributions and find that the beta-based specification provides a substantially better fit than gamma or inverse Gaussian alternatives, while yielding von Bertalanffy-type population summaries compatible with current stock-assessment practice. A simulation study shows that the proposed estimators recover key growth characteristics and reproduce the observed stepwise trajectories under realistic sample sizes. By embedding biologically constrained increments and intermoult timing within a stochastic jump-process framework, the model provides a flexible tool for analysing discontinuous growth in crustaceans and related biological systems, and illustrates how modern jump-process methods can inform fisheries management.

REFERENCES

BAZHBA, M., BLANCHET, J., RHEE, C.-H. and ZWART, B. (2020). Sample path large deviations for Lévy processes and random walks with Weibull increments. *Ann. Appl. Probab.* **30** 2695–2739. MR4187125 https://doi.org/10.1214/20-AAP1570

BERTOIN, J. (1996). *Lévy Processes*. *Cambridge Tracts in Mathematics* **121**. Cambridge Univ. Press, Cambridge. MR1406564

CHANG, Y. J., SUN, C. L., CHEN, Y. and YEH, S. Z. (2012). Modelling the growth of crustacean species. *Rev. Fish Biol. Fisher.* **22** 157–187.

CONT, R. and TANKOV, P. (2004). *Financial Modelling with Jump Processes*. *Chapman & Hall/CRC Financial Mathematics Series*. CRC Press/CRC, Boca Raton, FL. MR2042661

DENNIS, D. M., SKEWES, T. D. and PITCHER, C. R. (1997). Habitat use and growth of juvenile ornate rock lobsters, *Panulirus ornatus* (Fabricius, 1798), in Torres Strait, Australia. *Mar. Freshw. Res.* **48** 663–670.

DI CRESCENZO, A., MUSTO, S., PARAGGIO, P. and TORRES-RUIZ, F. (2025). Special lognormal diffusion processes with binomial random catastrophes and applications to economic data. *Appl. Math. Model.* **145** Paper No. 116146. MR4896293 https://doi.org/10.1016/j.apm.2025.116146

FABENS, A. J. (1965). Properties and fitting of the von Bertalanffy growth curve. *Growth* **29** 265–289.

FOO, C. H. (2020). Parameter estimation of laboratory-reared *Panulirus ornatus*. *SN Appl. Sci.* **2** 1–6.

FOO, C. H. (2023). Modeling discontinuous growth in reared *Panulirus ornatus*: A generalized additive model and Cox proportional hazard model approach. *Math. Biosci. Eng.* **20** 14487–14501. MR4614461 https://doi.org/10.3934/mbe.2023648

Key words and phrases. Stochastic growth modelling, Lévy subordinator, jump processes, convolution likelihood, moult increment, intermoult period, crustacean growth, fisheries stock assessment.

GARCIA-MARCH, J. R., GARCIA-CARRASCOSA, A. M., CANTERO, A. L. P. and WANG, Y. G. (2007). Population structure, mortality and growth of Pinna nobilis Linnaeus, 1758 (Mollusca, Bivalvia) at different depths in Moraira bay (Alicante, Western Mediterranean). *Mar. Biol.* **150** 861–871.

GREVEN, S. and KNEIB, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* **97** 773–789. MR2746151 https://doi.org/10.1093/biomet/asq042

GUDMUNDSSON, G. (2005). Stochastic growth. *Can. J. Fish. Aquat. Sci.* **62** 1746–1755.

HART, D. R. and CHUTE, A. S. (2009). Estimating von Bertalanffy growth parameters from growth increment data using a linear mixed-effects model, with an application to the sea scallop *Placopecten magellanicus*. *ICES J. Mar. Sci.* **66** 2165–2175.

HOENIG, J. M. and RESTREPO, V. R. (1989). Estimating the intermolt periods in asynchronously molting crustacean populations. *Biometrics* **45** 71–82.

JACKSON, C. J. and WANG, Y.-G. (1998). Modelling growth rate of Penaeus monodon Fabricius in intensively managed ponds: Effects of temperature, pond age and stocking density. *Aquac. Res.* **29** 27–36.

KIMURA, D. K. (1980). Likelihood methods for the von Bertalanffy growth curve. *Fish. Bull.* **77** 765–776.

LASLETT, G. M., EVESON, J. P. and POLACHECK, T. (2004). Fitting growth models to length frequency data. *ICES J. Mar. Sci.* **61** 218–230.

LV, Q. and PITCHFORD, J. W. (2007). Stochastic von Bertalanffy models, with applications to fish recruitment. *J. Theoret. Biol.* **244** 640–655. MR2306354 https://doi.org/10.1016/j.jtbi.2006.09.009

MILLAR, R. B. and HOENIG, J. M. (1997). A generalized model for estimating intermolt periods of asynchronously molting insects and crustacea from field or laboratory data. *J. Agric. Biol. Environ. Stat.* **2** 389–402. MR1817046 https://doi.org/10.2307/1400510

MYKLES, D. L. (2011). Ecdysteroid metabolism in crustaceans. *J. Steroid Biochem. Mol. Biol.* **127** 196–203.

RAABE, D., SACHS, C. and TRIGUERO, P. R. (2005). The crustacean exoskeleton as an example of a structurally and mechanically graded biological nanocomposite material. *Acta Mater.* **53** 4281–4292.

ROMÁN-ROMÁN, P., ROMERO, D. and TORRES-RUIZ, F. (2010). A diffusion process to model generalized von Bertalanffy growth patterns: Fitting to real data. *J. Theoret. Biol.* **263** 59–69. MR2980719 https://doi.org/10.1016/j.jtbi.2009.12.009

RUSSO, T., BALDI, P., PARISI, A., MAGNIFICO, G., MARIANI, S. and CATAUDELLA, S. (2009a). Lévy processes and stochastic von Bertalanffy models of growth, with application to fish population analysis. *J. Theoret. Biol.* **258** 521–529. MR2973261 https://doi.org/10.1016/j.jtbi.2009.01.033

RUSSO, T., MARIANI, S., BALDI, P., PARISI, A., MAGNIFICO, G., WORSØE, C. and CATAUDELLA, S. (2009b). Progress in modelling herring populations: An individual-based model of growth. *ICES J. Mar. Sci.* **66** 1718–1725.

SAPUTRA, I. and PRIYAMBODO, B. (2023). The development of lobster puerulus (*Panulirus ornatus* and *P. homarus*) in captivity environment. *Indones. Aquacult. J.* **18** 167–177.

SATO, K. (1999). *Lévy Processes and Infinitely Divisible Distributions. Cambridge Studies in Advanced Mathematics* **68**. Cambridge Univ. Press, Cambridge. MR1739520

SUDEWI, N. A. G., SLAMET, B. and HARYANTI (2023). Growth at molt of sub-adult scalloped spiny lobster Panulirus homarus reared in tanks. *IOP Conf. Ser. Earth Environ. Sci.* **1221** 1–8.

TOVAR-ÁVILA, J., TROYNIKOV, V. S., WALKER, T. I. and DAY, R. W. (2009). Use of stochastic models to estimate the growth of the Port Jackson shark, *Heterodontus portusjacksoni*, off eastern Victoria, Australia. *Fish. Res.* **95** 230–235.

VEILLETTE, M. and TAQQU, M. S. (2010). Numerical computation of first passage times of increasing Lévy processes. *Methodol. Comput. Appl. Probab.* **12** 695–729. MR2726540 https://doi.org/10.1007/s11009-009-9158-y

VON BERTALANFFY, L. (1938). A quantitative theory of organic growth (inquiries on growth laws. II). *Hum. Biol.* **10** 181–213.

WANG, Y.-G. (1998). An improved Fabens method for estimation of growth parameters in the vonBertalanffy model with individual asymptotes. *Can. J. Fish. Aquat. Sci.* **55** 397–400.

WANG, Y.-G. (1999). Estimating equations for parameters in stochastic growth models from tag-recapture data. *Biometrics* **55** 900–903.

WANG, Y.-G. and ELLIS, N. (2005). Maximum likelihood estimation of mortality and growth with individual variability from multiple length-frequency data. *Fish. Bull.* **103** 380–391.

WANG, Y.-G. and FOO, C. H. (2026). Supplement to "Lévy processes for jumping growth of spiny lobsters, *Panulirus ornatus*." https://doi.org/10.1214/26-AOAS2188SUPP

WANG, Y.-G. and THOMAS, M. R. (1995). Accounting for individual variability in the von Bertalanffy growth-model. *Can. J. Fish. Aquat. Sci.* **52** 1368–1375.

WANG, Y.-G., THOMAS, M. R. and SOMERS, I. F. (1995). A maximum-likelihood approach for estimating growth from tag-recapture data. *Can. J. Fish. Aquat. Sci.* **52** 252–259.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1227–1246
https://doi.org/10.1214/26-AOAS2176

CONFIDENCE INTERVALS FOR RATE ESTIMATION WITH IMPORTANCE SAMPLING IN AUTONOMOUS VEHICLE EVALUATION

BY AIYOU CHEN[a], RUIXUAN RACHEL ZHOU[b], JOSEPH J. LEE[c], NICHOLAS CHAMANDY[d] AND HENNING HOHNHOLD[e]

Waymo LLC, [a]*aiyouchen@waymo.com,* [b]*rachelzhou@waymo.com,* [c]*josephjlee@waymo.com,* [d]*chamandy@waymo.com,* [e]*henningh@waymo.com*

Accounting for both rare events and complex sampling presents challenges when quantifying uncertainty for rate estimation in autonomous vehicle performance evaluation. In this paper we introduce a statistical formulation of this problem and develop a unified compound Poisson model framework for unbiased rate estimation through the Horvitz–Thompson estimator. Though asymptotic theory for the model is available, the inference of confidence intervals (CIs) in the presence of rare events requires new investigation. We also advocate for a new monotonicity criterion for rate CIs—summing the rates of disjoint types of events should produce not only a higher point estimate but also higher confidence bounds than for the individual rates—that facilitates interpretability in real applications. We propose a novel *exponential bootstrap* (EB) method for CI construction based on a fiducial argument; it satisfies the monotonicity property, while novel extensions of some existing methods do not. Comprehensive numerical studies show that EB performs well for a wide range of settings relevant to our applications. Fast implementation of EB based on saddlepoint approximation is also developed, which may be of independent interest.

REFERENCES

AGARWAL, A., XIAO, M., BARTER, R., RONEN, O., FAN, B. and YU, B. (2025). Pcs-uq: Uncertainty quantification via the predictability-computability-stability framework. arXiv preprint. Available at arXiv:2505.08784.

BICKEL, P. J. and DOKSUM, K. A. (2015). *Mathematical Statistics*: *Basic Ideas and Selected Topics*, *Volumes I*, 2nd ed. CRC Press, Boca Raton, FL.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. *Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

BOHM, G. and ZECH, G. (2014). Statistics of weighted Poisson events and its applications. *Nucl*. *Instrum*. *Methods Phys*. *Res*., *Sect*. *A*, *Accel*. *Spectrom*. *Detect*. *Assoc*. *Equip*. **748** 1–6.

ČEKANAVIČIUS, V. and NOVAK, S. Y. (2025). *Compound Poisson Approximation*. CRC Press, Boca Raton, FL. MR4935814

CHAMANDY, N., MURALIDHARAN, O., NAJMI, A. and NAIDU, S. (2012). Estimating uncertainty for massive data streams. Technical Report, Google Inc. Available at https://research.google/pubs/estimating-uncertainty-for-massive-data-streams.

CHEN, A., ZHOU, R. R., LEE, J. J., CHAMANDY, N. and HOHNHOLD, H. (2026). Supplement to "Confidence intervals for rate estimation with importance sampling in autonomous vehicle evaluation." https://doi.org/10.1214/26-AOAS2176SUPP

CHEN, Y.-H., SCANLON, J. M., KUSANO, K. D., MCMURRY, T. L. and VICTOR, T. (2024). Dynamic benchmarks: Spatial and temporal alignment for ADS performance evaluation. arXiv preprint. Available at arXiv:2410.08903.

DOBSON, A. J., KUULASMAA, K., EBERLE, E. and SCHERER, J. (1991). Confidence intervals for weighted sums of Poisson parameters. *Stat*. *Med*. **10** 457–462.

EFRON, B. and LEPAGE, R. (1992). Introduction to bootstrap. In *Exploring the Limits of Bootstrap* (*East Lansing*, *MI*, 1990). *Wiley Ser*. *Probab*. *Math*. *Statist*. *Probab*. *Math*. *Statist*. 3–10. Wiley, New York. MR1197776

Key words and phrases. Autonomous vehicles, confidence interval, compound Poisson, rare events, importance sampling.

FAY, M. P. and FEUER, E. J. (1997). Confidence intervals for directly standardized rates: A method based on the gamma distribution. *Stat. Med.* **16** 791–801.
FAY, M. P. and KIM, S. (2017). Confidence intervals for directly standardized rates using mid-p gamma intervals. *Biom. J.* **59** 377–387. MR3623347 https://doi.org/10.1002/bimj.201600111
FELLER, W. (2008). *An Introduction to Probability Theory and Its Applications*, *Vol* 2. Wiley, New York. MR0038583
FISHER, R. A. (1935). The fiducial argument in statistical inference. *Ann. Eugen.* **6** 391–398.
GARWOOD, F. (1936). Fiducial limits for the Poisson distribution. *Biometrika* **28** 437–442.
HANNIG, J., IYER, H., LAI, R. C. S. and LEE, T. C. M. (2016). Generalized fiducial inference: A review and new results. *J. Amer. Statist. Assoc.* **111** 1346–1361. MR3561954 https://doi.org/10.1080/01621459.2016.1165102
HESTERBERG, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37** 185–194.
HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460
JIANG, C. M., BAI, Y., CORNMAN, A., DAVIS, C., HUANG, X., JEON, H., KULSHRESTHA, S., LAMBERT, J. W., LI, S. et al. (2024). SceneDiffuser: Efficient and controllable driving simulation initialization and rollout. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*.
KABAILA, P. and LLOYD, C. J. (2006). Improved Buehler limits based on refined designated statistics. *J. Statist. Plann. Inference* **136** 3145–3155. MR2256221 https://doi.org/10.1016/j.jspi.2004.11.011
KALRA, N. and PADDOCK, S. M. (2016). Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transp. Res., Part A Policy Pract.* **94** 182–193.
KEGLER, S. R. (2007). Applying the compound Poisson process model to the reporting of injury-related mortality rates. *Epidemiol. Perspect. Innov.* **4** 1–9.
KUSANO, K. D., SCANLON, J. M., CHEN, Y.-H., MCMURRY, T. L., CHEN, R., GODE, T. and VICTOR, T. (2024). Comparison of Waymo rider-only crash data to human benchmarks at 7.1 million miles. *Traffic Inj. Prev.* **25** S66–S77.
LI, Z., WANG, X. and ZHOU, W. (2012). Empirical likelihood for compound Poisson processes. *Aust. N. Z. J. Stat.* **54** 463–474. MR3018628 https://doi.org/10.1111/j.1467-842X.2012.00678.x
LILLO, L. D., GODE, T., ZHOU, X., ATZEI, M., CHEN, R. and VICTOR, T. (2024a). Comparative safety performance of autonomous-and human drivers: A real-world case study of the Waymo driver. *Heliyon* **10**.
LILLO, L. D., GODE, T., ZHOU, X., SCANLON, J., CHEN, R. and VICTOR, T. (2024b). Do Autonomous Vehicles Outperform Latest-Generation Human-Driven Vehicles? A Comparison to Waymo's Auto Liability Insurance Claims at 25 Million Miles. Technical Report, Waymo LLC. Available at https://waymo.com/research.
LIU, D., LIU, R. Y. and XIE, M. (2014). Exact meta-analysis approach for discrete data and its application to 2×2 tables with rare events. *J. Amer. Statist. Assoc.* **109** 1450–1465. MR3293603 https://doi.org/10.1080/01621459.2014.946318
LIU, Y. (2019). Estimating the prevalence of rare events—theory and practice. Technical Report, Google Inc. Available at https://www.unofficialgoogledatascience.com/2019/08/estimating-prevalence-of-rare-events.html.
MAHJOURIAN, R., MU, R., LIKHOSHERSTOV, V., MOUGIN, P., HUANG, X., MESSIAS, J. V. and WHITESON, S. (2024). UniGen: Unified modeling of initial agent states and trajectories for generating autonomous driving scenarios. In *IEEE International Conference on Robotics and Automation*.
MARTIN, R. and LIU, C. (2015). Marginal inferential models: Prior-free probabilistic inference on interest parameters. *J. Amer. Statist. Assoc.* **110** 1621–1631. MR3449059 https://doi.org/10.1080/01621459.2014.985827
MCCORMICK, T. C. (1937). Sampling theory in sociological research. *Soc. Forces* **16** 67.
NG, H. K. T., FILARDO, G. and ZHENG, G. (2008). Confidence interval estimating procedures for standardized incidence rates. *Comput. Statist. Data Anal.* **52** 3501–3516. MR2427360 https://doi.org/10.1016/j.csda.2007.11.004
OWEN, A. and ZHOU, Y. (2000). Safe and effective importance sampling. *J. Amer. Statist. Assoc.* **95** 135–143. MR1803146 https://doi.org/10.2307/2669533
OWEN, A. B. (2001). *Empirical Likelihood*. CRC Press, Boca Raton.
SEN, B., WALKER, M. and WOODROOFE, M. (2009). On the unified method with nuisance parameters. *Statist. Sinica* **19** 301–314. MR2487891
SINHA, A., NIKDEL, P., PAUL, S. and WHITESON, S. (2025). Rate-informed discovery via Bayesian adaptive multifidelity sampling. In *Conference on Robot Learning* 2579–2598. PMLR.
SUGIYAMA, M., KRAULEDAT, M. and MÜLLER, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.* **8**.
SWIFT, M. B. (2010). A simulation study comparing methods for calculating confidence intervals for directly standardized rates. *Comput. Statist. Data Anal.* **54** 1103–1108. MR2580941 https://doi.org/10.1016/j.csda.2009.10.021

TERRES, M. A., CHEN, A., ZHOU, R. and MCLEOD, C. M. (2023). Behavioral event detection and rate estimation for autonomous vehicle evaluation. *Appl. Stoch. Models Bus. Ind.* **39** 662–683. MR4662209 https://doi.org/10.1002/asmb.2769

TIWARI, R. C., CLEGG, L. X. and ZOU, Z. (2006). Efficient interval estimation for age-adjusted cancer rates. *Stat. Methods Med. Res.* **15** 547–569. MR2312664 https://doi.org/10.1177/0962280206070621

WEBB, N., SMITH, D., LUDWICK, C., VICTOR, T. W., HOMMES, Q., FAVARO, F., IVANOV, G. and DANIEL, T. (2020). Waymo's Safety Methodologies and Safety Readiness Determinations. Technical Report, Waymo LLC. Available at https://www.waymo.com/safety.

XIE, M. and WANG, P. (2024). Repro samples method for a performance guaranteed inference in general and irregular inference problems. arXiv preprint. Available at arXiv:2402.15004.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1247–1266
https://doi.org/10.1214/26-AOAS2167

SPECTRAL-STIMULUS INFORMATION FOR SELF-SUPERVISED STIMULUS ENCODING

BY JARED DEIGHTON[1,2,b], WYATT MACKEY[3,c], IOANNIS SCHIZAS[3,d], DAVID L. BOOTHE[3,e] AND VASILEIOS MAROULAS[1,a]

[1]*Department of Mathematics, University of Tennessee,* [a]*vmaroula@utk.edu*

[2]*Department of Computer, Data, & Mathematical Sciences, Simmons University,* [b]*jared.deighton@simmons.edu*

[3]*The U.S. Army Combat Capabilities Development Command Army Research Laboratory,* [c]*wyatt.t.mackey.civ@army.mil,* [d]*ioannis.d.schizas.civ@army.mil,* [e]*david.l.boothe7.civ@army.mil*

Understanding how neural populations efficiently encode stimuli is a fundamental challenge in computational neuroscience. Existing rate-based information measures primarily assess single-neuron encoding, limiting insights into population-level representations, while the role of correlation in neural coding remains a subject of considerable debate. To address this, we introduce novel, correlation-aware information-theoretic measures that quantify the encoding efficiency of multiple neurons, including the joint stimulus information rate for neuron pairs and the spectral-stimulus information for arbitrarily sized populations. The spectral-stimulus information, defined as the leading eigenvalue of the stimulus information matrix, is maximized when neurons exhibit localized, nonoverlapping firing fields. We apply these measures to the domain of spatial navigation, where specialized neurons, such as place cells, grid cells, and head direction cells, encode position and orientation from self-motion and environmental cues. Analyzing neural recordings from mice and monkeys, we elucidate differences in encoding efficiency across neuronal pairs and populations. We then demonstrate that these measures can be used to train recurrent neural networks (RNNs) via self-supervised learning, leading to the emergence of place cells and head direction cells. Our findings provide a principled population-level framework for understanding stimulus encoding, with broad implications for neuroscience and the optimization of artificial navigation systems.

REFERENCES

ALME, C. B., MIAO, C., JEZEK, K., TREVES, A., MOSER, E. I. and MOSER, M.-B. (2014). Place cells in the hippocampus: Eleven maps for eleven rooms. *Proc. Natl. Acad. Sci. USA* **111** 18428–18435.

AMARI, S. and NAKAHARA, H. (2006). Correlation and independence in the neural code. *Neural Comput.* **18** 1259–1267. MR2218836 https://doi.org/10.1162/neco.2006.18.6.1259

ANGELAKI, D. E. and LAURENS, J. (2020). The head direction cell network: Attractor dynamics, integration within the navigation system, and three-dimensional properties. *Curr. Opin. Neurobiol.* **60** 136–144.

ARONOV, D., NEVERS, R. and TANK, D. W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature* **543** 719–722.

BANINO, A., BARRY, C., URIA, B., BLUNDELL, C., LILLICRAP, T., MIROWSKI, P., PRITZEL, A., CHADWICK, M. J., DEGRIS, T. et al. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature* **557** 429–433.

BENNA, M. K. and FUSI, S. (2021). Place cells may simply be memory cells: Memory compression leads to spatial tuning and history dependence. *Proc. Natl. Acad. Sci. USA* **118**. e2018422118.

BRENNER, N., STRONG, S. P., KOBERLE, R., BIALEK, W. and STEVENINCK, R. R. D. R. V. (2000). Synergy in a neural code. *Neural Comput.* **12** 1531–1552.

BURAK, Y. and FIETE, I. R. (2009). Accurate path integration in continuous attractor network models of grid cells. *PLoS Comput. Biol.* **5** e1000291, 16. MR2496586 https://doi.org/10.1371/journal.pcbi.1000291

BUTTS, D. A. (2003). How much information is associated with a particular stimulus? *Netw. Comput. Neural Syst.* **14** 177.

Key words and phrases. Information theory, recurrent neural networks, place cells, neural coding.

BUTTS, D. A. and GOLDMAN, M. S. (2006). Tuning curves, neuronal variability, and sensory coding. *PLoS Biol.* **4** e92.

CHECHIK, G., GLOBERSON, A., ANDERSON, M., YOUNG, E., NELKEN, I. and TISHBY, N. (2001). Group redundancy measures reveal redundancy reduction in the auditory pathway. *Adv. Neural Inf. Process. Syst.* **14**.

CHEN, G., KING, J. A., BURGESS, N. and O'KEEFE, J. (2013). How vision and movement combine in the hippocampal place code. *Proc. Natl. Acad. Sci. USA* **110** 378–383.

CLAUW, K., STRAMAGLIA, S. and MARINAZZO, D. (2024). Information-theoretic progress measures reveal grokking is an emergent phase transition. arXiv preprint. Available at arXiv:2408.08944.

COUEY, J. J., WITOELAR, A., ZHANG, S.-J., ZHENG, K., YE, J., DUNN, B., CZAJKOWSKI, R., MOSER, M.-B., MOSER, E. I. et al. (2013). Recurrent inhibitory circuitry as a mechanism for grid formation. *Nat. Neurosci.* **16** 318–324.

CUEVA, C. J. and WEI, X.-X. (2018). Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. arXiv preprint. Available at arXiv:1803.07770.

DAYAN, P. and ABBOTT, L. F. (2005). *Theoretical Neuroscience*: *Computational and Mathematical Modeling of Neural Systems*. *Computational Neuroscience*. MIT Press, Cambridge, MA. MR1985615

DEIGHTON, J., MACKEY, W., SCHIZAS, I., BOOTHE, D. L. and MAROULAS, V. (2026). Supplement to "Spectral-stimulus information for self-supervised stimulus encoding." https://doi.org/10.1214/26-AOAS2167SUPPA, https://doi.org/10.1214/26-AOAS2167SUPPB, https://doi.org/10.1214/26-AOAS2167SUPPC

DORRELL, W., LATHAM, P. E., BEHRENS, T. E. and WHITTINGTON, J. C. (2022). Actionable neural representations: Grid cells from minimal constraints. arXiv preprint. Available at arXiv:2209.15563.

DUPRET, D., O'NEILL, J., PLEYDELL-BOUVERIE, B. and CSICSVARI, J. (2010). The reorganization and reactivation of hippocampal maps predict spatial memory performance. *Nat. Neurosci.* **13** 995–1002.

ELIAV, T., MAIMON, S. R., ALJADEFF, J., TSODYKS, M., GINOSAR, G., LAS, L. and ULANOVSKY, N. (2021). Multiscale representation of very large environments in the hippocampus of flying bats. *Science* **372**. eabg4020.

ELMAN, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Mach. Learn.* **7** 195–225.

FENTON, A. A., KAO, H.-Y., NEYMOTIN, S. A., OLYPHER, A., VAYNTRUB, Y., LYTTON, W. W. and LUDVIG, N. (2008). Unmasking the CA1 ensemble place code by exposures to small and large environments: More place cells and multiple, irregularly arranged, and expanded place fields in the larger space. *J. Neurosci.* **28** 11250–11262.

FINKELSTEIN, A., DERDIKMAN, D., RUBIN, A., FOERSTER, J. N., LAS, L. and ULANOVSKY, N. (2015). Three-dimensional head-direction coding in the bat brain. *Nature* **517** 159–164.

FRANK, L. M., BROWN, E. N. and WILSON, M. (2000). Trajectory encoding in the hippocampus and entorhinal cortex. *Neuron* **27** 169–178.

FU, H., RODRIGUEZ, G. A., HERMAN, M., EMRANI, S., NAHMANI, E., BARRETT, G., FIGUEROA, H. Y., GOLDBERG, E., HUSSAINI, S. A. et al. (2017). Tau pathology induces excitatory neuron loss, grid cell dysfunction, and spatial memory deficits reminiscent of early Alzheimer's disease. *Neuron* **93** 533–541.

FUHS, M. C. and TOURETZKY, D. S. (2006). A spin glass model of path integration in rat medial entorhinal cortex. *J. Neurosci.* **26** 4266–4276.

GANTMACHER, F. R. (2000). *The Theory of Matrices* 131. Amer. Math. Soc., Providence.

GARDNER, R. J., HERMANSEN, E., PACHITARIU, M., BURAK, Y., BAAS, N. A., DUNN, B. A., MOSER, M.-B. and MOSER, E. I. (2022). Toroidal topology of population activity in grid cells. *Nature* **602** 123–128.

GAUTHIER, J. L. and TANK, D. W. (2018). A dedicated population for reward coding in the hippocampus. *Neuron* **99** 179–193.

GUANELLA, A., KIPER, D. and VERSCHURE, P. (2007). A model of grid cells based on a twisted torus topology. *Int. J. Neural Syst.* **17** 231–240.

HAFTING, T., FYHN, M., MOLDEN, S., MOSER, M.-B. and MOSER, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature* **436** 801–806.

HARLAND, B., CONTRERAS, M., SOUDER, M. and FELLOUS, J.-M. (2021). Dorsal CA1 hippocampal place cells form a multi-scale representation of megaspace. *Curr. Biol.* **31** 2178–2190.

HAYMAN, R., VERRIOTIS, M. A., JOVALEKIC, A., FENTON, A. A. and JEFFERY, K. J. (2011). Anisotropic encoding of three-dimensional space by place cells and grid cells. *Nat. Neurosci.* **14** 1182–1188.

HAZAMA, Y. and TAMURA, R. (2019). Data on the activity of place cells in the hippocampal CA1 subfield of a monkey performing a shuttling task. *Data Brief* **26** 104467.

ISSA, J. B. and ZHANG, K. (2012). Universal conditions for exact path integration in neural systems. *Proc. Natl. Acad. Sci. USA* **109** 6716–6720.

KAYSER, C., LOGOTHETIS, N. K. and PANZERI, S. (2010). Millisecond encoding precision of auditory cortex neurons. *Proc. Natl. Acad. Sci. USA* **107** 16976–16981.

KEINATH, A. T., JULIAN, J. B., EPSTEIN, R. A. and MUZZIO, I. A. (2017). Environmental geometry aligns the hippocampal map during spatial reorientation. *Curr. Biol.* **27** 309–317.

KROPFF, E., CARMICHAEL, J. E., MOSER, M.-B. and MOSER, E. I. (2015). Speed cells in the medial entorhinal cortex. *Nature* **523** 419–424.

LATHAM, P. E. and ROUDI, Y. (2013). Role of correlations in population coding. In *Principles of Neural Coding* 121–138.

LEVER, C., BURTON, S., JEEWAJEE, A., O'KEEFE, J. and BURGESS, N. (2009). Boundary vector cells in the subiculum of the hippocampal formation. *J. Neurosci.* **29** 9771–9777.

LEVER, C., WILLS, T., CACUCCI, F., BURGESS, N. and O'KEEFE, J. (2002). Long-term plasticity in hippocampal place-cell representation of environmental geometry. *Nature* **416** 90–94.

LEVY, S. J. and HASSELMO, M. E. (2023). Hippocampal remapping induced by new behavior is mediated by spatial context. bioRxiv 2023-02.

MAINALI, N., DA SILVEIRA, R. A. and BURAK, Y. (2025). Universal statistics of hippocampal place fields across species and dimensionalities. *Neuron* **113** 1110–1120.

MALERBA, S. B., PIEROPAN, M., BURAK, Y. and DA SILVEIRA, R. A. (2025). Random compressed coding with neurons. *Cell Rep.* **44**.

MARKUS, E. J., BARNES, C. A., MCNAUGHTON, B. L., GLADDEN, V. L. and SKAGGS, W. E. (1994). Spatial information content and reliability of hippocampal CA1 neurons: Effects of visual input. *Hippocampus* **4** 410–421.

MCNAUGHTON, B. L., BATTAGLIA, F. P., JENSEN, O., MOSER, E. I. and MOSER, M.-B. (2006). Path integration and the neural basis of the'cognitive map'. *Nat. Rev. Neurosci.* **7** 663–678.

MITCHELL, E. C., STORY, B., BOOTHE, D., FRANASZCZUK, P. J. and MAROULAS, V. (2024). A topological deep learning framework for neural spike decoding. *Biophys. J.*

NASRIN, F., OBALLE, C., BOOTHE, D. and MAROULAS, V. (2019). Bayesian topological learning for brain state classification. In 2019 18*th IEEE International Conference on Machine Learning and Applications* (*ICMLA*) 1247–1252. IEEE Press, New York.

NIEH, E. H., SCHOTTDORF, M., FREEMAN, N. W., LOW, R. J., LEWALLEN, S., KOAY, S. A., PINTO, L., GAUTHIER, J. L., BRODY, C. D. et al. (2021). Geometry of abstract learned knowledge in the hippocampus. *Nature* **595** 80–84.

O'KEEFE, J. (1976). Place units in the hippocampus of the freely moving rat. *Exp. Neurol.* **51** 78–109.

O'KEEFE, J. and DOSTROVSKY, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.*

O'KEEFE, J. and NADEL, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford Univ. Press, London.

ORMOND, J. and O'KEEFE, J. (2022). Hippocampal place cells have goal-oriented vector fields during navigation. *Nature* **607** 741–746.

PETTERSEN, M., ROGGE, F. and LEPPERØD, M. (2024). Learning place cell representations and context-dependent remapping. *Adv. Neural Inf. Process. Syst.* **37** 244–269.

PETTERSEN, M., SCHØYEN, V. S., ØSTBY, M. D., MALTHE-SØRENSSEN, A. and LEPPERØD, M. E. (2024). Self-supervised grid cells without path integration. bioRxiv 2024-05.

PEYRACHE, A., LACROIX, M. M., PETERSEN, P. C. and BUZSÁKI, G. (2015). Internally organized mechanisms of the head direction sense. *Nat. Neurosci.* **18** 569–575.

QUIAN QUIROGA, R. and PANZERI, S. (2009). Extracting information from neuronal populations: Information theory and decoding approaches. *Nat. Rev. Neurosci.* **10** 173–185.

RAJU, R. V., GUNTUPALLI, J. S., ZHOU, G., WENDELKEN, C., LÁZARO-GREDILLA, M. and GEORGE, D. (2024). Space is a latent sequence: A theory of the hippocampus. *Sci. Adv.* **10** eadm8470.

REDMAN, W., ACOSTA, F., ACOSTA-MENDOZA, S. and MIOLANE, N. (2024). Not so griddy: Internal representations of RNNs path integrating more than one agent. *Adv. Neural Inf. Process. Syst.* **37** 22657–22689.

RICH, P. D., LIAW, H.-P. and LEE, A. K. (2014). Large environments reveal the statistical structure governing hippocampal representations. *Science* **345** 814–817.

SANDERS, H., JI, D., SASAKI, T., LEUTGEB, J. K., WILSON, M. A. and LISMAN, J. E. (2019). Temporal coding and rate remapping: Representation of nonspatial information in the hippocampus. *Hippocampus* **29** 111–127.

SAVARESE, P., KIM, S. S., MAIRE, M., SHAKHNAROVICH, G. and MCALLESTER, D. (2021). Information-theoretic segmentation by inpainting error maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 4029–4039.

SAVE, E., CRESSANT, A., THINUS-BLANC, C. and POUCET, B. (1998). Spatial firing of hippocampal place cells in blind rats. *J. Neurosci.* **18** 1818–1826.

SCHAEFFER, R., KHONA, M. and FIETE, I. (2022). No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. *Adv. Neural Inf. Process. Syst.* **35** 16052–16067.

SCHAEFFER, R., KHONA, M., MA, T., EYZAGUIRRE, C., KOYEJO, S. and FIETE, I. (2024). Self-supervised learning of representations for space generates multi-modular grid cells. *Adv. Neural Inf. Process. Syst.* **36**.

SCHNEIDMAN, E., BIALEK, W. and BERRY, M. J. (2003). Synergy, redundancy, and independence in population codes. *J. Neurosci.* **23** 11539–11553.

SENGUPTA, A., PEHLEVAN, C., TEPPER, M., GENKIN, A. and CHKLOVSKII, D. (2018). Manifold-tiling localized receptive fields are optimal in similarity-preseUniversal conditions for exact path integration in neural systemsrving neural networks. *Adv. Neural Inf. Process. Syst.* **31**.

SERIÈS, P., LATHAM, P. E. and POUGET, A. (2004). Tuning curve sharpening for orientation selectivity: Coding efficiency and the impact of correlations. *Nat. Neurosci.* **7** 1129–1135.

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27** 379–423. MR0026286 https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

SKAGGS, W., MCNAUGHTON, B. and GOTHARD, K. (1992). An information-theoretic approach to deciphering the hippocampal code. *Adv. Neural Inf. Process. Syst.* **5**.

SORSCHER, B., MEL, G., GANGULI, S. and OCKO, S. (2019). A unified theory for the origin of grid cells through the lens of pattern formation. *Adv. Neural Inf. Process. Syst.* **32**.

SORSCHER, B., MEL, G. C., OCKO, S. A., GIOCOMO, L. M. and GANGULI, S. (2023). A unified theory for the computational and mechanistic origins of grid cells. *Neuron* **111** 121–137.

SOUZA, B. C., PAVÃO, R., BELCHIOR, H. and TORT, A. B. (2018). On information metrics for spatial coding. *Neuroscience* **375** 62–73.

SQUIRE, L. R., STARK, C. E. and CLARK, R. E. (2004). The medial temporal lobe. *Annu. Rev. Neurosci.* **27** 279–306.

STACHENFELD, K. L., BOTVINICK, M. M. and GERSHMAN, S. J. (2017). The hippocampus as a predictive map. *Nat. Neurosci.* **20** 1643–1653.

STENSOLA, T. and MOSER, E. I. (2016). Grid cells and spatial maps in entorhinal cortex and hippocampus. In *Micro-, Meso-and Macro-Dynamics of the Brain* 59–80.

SUN, W., WINNUBST, J., NATRAJAN, M., LAI, C., KAJIKAWA, K., BAST, A., MICHAELOS, M., GATTONI, R., STRINGER, C. et al. (2025). Learning produces an orthogonalized state machine in the hippocampus. *Nature* 1–11.

TAUBE, J. S., MULLER, R. U. and RANCK, J. B. (1990). Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *J. Neurosci.* **10** 420–435.

TIMME, N., ALFORD, W., FLECKER, B. and BEGGS, J. M. (2014). Synergy, redundancy, and multivariate information measures: An experimentalist's perspective. *J. Comput. Neurosci.* **36** 119–140. MR3176934 https://doi.org/10.1007/s10827-013-0458-4

WANG, X., LU, T., SNIDER, R. K. and LIANG, L. (2005). Sustained firing in auditory cortex evoked by preferred stimuli. *Nature* **435** 341–346.

WANG, Y., XU, X. and WANG, R. (2019). The place cell activity is information-efficient constrained by energy. *Neural Netw.* **116** 110–118.

WANG, Z., DI TULLIO, R. W., ROOKE, S. and BALASUBRAMANIAN, V. (2024). Time makes space: Emergence of place fields in networks encoding temporally continuous sensory experiences. bioRxiv.

WHITTINGTON, J. C., MULLER, T. H., MARK, S., CHEN, G., BARRY, C., BURGESS, N. and BEHRENS, T. E. (2020). The Tolman–Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell* **183** 1249–1263.

WILLIAMS, P. L. and BEER, R. D. (2010). Nonnegative decomposition of multivariate information. arXiv preprint. Available at arXiv:1004.2515.

YODER, R. M., PECK, J. R. and TAUBE, J. S. (2015). Visual landmark information gains control of the head direction signal at the lateral mammillary nuclei. *J. Neurosci.* **35** 1354–1367.

YU, C., BEHRENS, T. E. and BURGESS, N. (2020). Prediction and generalisation over directed actions by grid cells. arXiv preprint. Available at arXiv:2006.03355.

ZBONTAR, J., JING, L., MISRA, I., LECUN, Y. and DENY, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning* 12310–12320. PMLR.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1267–1282
https://doi.org/10.1214/26-AOAS2198

GENERATIVE SCORE INFERENCE FOR MULTIMODAL DATA

BY XINYU TIAN[a] AND XIAOTONG SHEN[b]

School of Statistics, University of Minnesota, [a]*tianx@umn.edu,* [b]*xshen@umn.edu*

Accurate uncertainty quantification is crucial for making reliable decisions in various supervised learning scenarios, particularly when dealing with complex, multimodal data such as images and text. Current approaches often face notable limitations, including rigid assumptions and limited generalizability, constraining their effectiveness across diverse supervised learning tasks. To overcome these limitations, we introduce Generative Score Inference (GSI), a flexible inference framework capable of constructing statistically valid and informative prediction and confidence sets across a wide range of multimodal learning problems. GSI utilizes synthetic samples generated by deep generative models to approximate conditional score distributions, facilitating precise uncertainty quantification without imposing restrictive assumptions about the data or tasks. We empirically validate GSI's capabilities through two representative scenarios: hallucination detection in large language models and uncertainty estimation in image captioning. Our method achieves state-of-the-art performance in hallucination detection and robust predictive uncertainty in image captioning, and its performance is positively influenced by the quality of the underlying generative model. These findings underscore the potential of GSI as a versatile inference framework, significantly enhancing uncertainty quantification and trustworthiness in multimodal learning.

REFERENCES

AI@META (2024). Llama 3 Model Card.

ALAA, A. M., HUSSAIN, Z. and SONTAG, D. (2023). Conformalized unconditional quantile regression. In *International Conference on Artificial Intelligence and Statistics* 10690–10702. PMLR.

ANGELOPOULOS, A. N. and BATES, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint. Available at arXiv:2107.07511.

BALTRUŠAITIS, T., AHUJA, C. and MORENCY, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41** 423–443.

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, Methodol.* **57** 289–300. MR1325392

CHEN, T. and GUESTRIN, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining* 785–794.

CHUNG, Y., NEISWANGER, W., CHAR, I. and SCHNEIDER, J. (2021). Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. *Adv. Neural Inf. Process. Syst.* **34** 10971–10984.

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

DAI, B., SHEN, X. and WONG, W. (2022). Coupled generation. *J. Amer. Statist. Assoc.* **117** 1243–1253. MR4480709 https://doi.org/10.1080/01621459.2020.1844719

FARQUHAR, S., KOSSEN, J., KUHN, L. and GAL, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature* **630** 625–630.

FELDMAN, S., BATES, S. and ROMANO, Y. (2021). Improving conditional coverage via orthogonal quantile regression. *Adv. Neural Inf. Process. Syst.* **34** 2060–2071.

GIBNEY, E. (2022). Is AI fuelling a reproducibility crisis in science. *Nature* **608** 250–251.

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

GUI, Y., JIN, Y. and REN, Z. (2024). Conformal alignment: Knowing when to trust foundation models with guarantees. arXiv preprint. Available at arXiv:2405.10301.

Key words and phrases. Statistical uncertainty, multimodality, prediction sets, generative models, trustworthy.

GUO, C., PLEISS, G., SUN, Y. and WEINBERGER, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning* 1321–1330. PMLR.

HO, J., JAIN, A. and ABBEEL, P. (2020). Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33** 6840–6851.

JIN, Y. and CANDÈS, E. J. (2023). Selection by prediction with conformal p-values. *J. Mach. Learn. Res.* **24** Paper No. [244], 41. MR4633971

JUNG, C., NOAROV, G., RAMALINGAM, R. and ROTH, A. (2023). Batch multivalid conformal prediction. In *International Conference on Learning Representations* (*ICLR*).

LI, J., LI, D., XIONG, C. and HOI, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning* 12888–12900. PMLR.

LIU, Y., SHEN, R. and SHEN, X. (2024). Novel uncertainty quantification through perturbation-assisted sample synthesis. *IEEE Trans. Pattern Anal. Mach. Intell.* **46** 7813–7824.

MINDERER, M., DJOLONGA, J., ROMIJNDERS, R., HUBIS, F., ZHAI, X., HOULSBY, N., TRAN, D. and LUCIC, M. (2021). Revisiting the calibration of modern neural networks. *Adv. Neural Inf. Process. Syst.* **34** 15682–15694.

MIRZA, M. and OSINDERO, S. (2014). Conditional generative adversarial nets. arXiv preprint. Available at arXiv:1411.1784.

REZENDE, D. and MOHAMED, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning* 1530–1538. PMLR.

ROMANO, Y., PATTERSON, E. and CANDES, E. (2019). Conformalized quantile regression. *Adv. Neural Inf. Process. Syst.* **32**.

SHAFER, G. and VOVK, V. (2008). A tutorial on conformal prediction. *J. Mach. Learn. Res.* **9** 371–421. MR2417240

SOHN, K., LEE, H. and YAN, X. (2015). Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems* **28**.

SRIRAMANAN, G., BHARTI, S., SADASIVAN, V. S., SAHA, S., KATTAKINDA, P. and FEIZI, S. (2024). Llm-check: Investigating detection of hallucinations in large language models. *Adv. Neural Inf. Process. Syst.* **37** 34188–34216.

TIAN, X. and SHEN, X. (2026a). Supplement to "Generative score inference for multimodal data." https://doi.org/10.1214/26-AOAS2198SUPP

TIAN, X. and SHEN, X. (2026b). Enhancing accuracy in generative models via knowledge transfer. *J. Mach. Learn. Res.* To appear. arXiv preprint. Available at arXiv:2405.16837.

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic Learning in a Random World*, *Vol.* 29. Springer, New York. MR2161220

WINKLER, C., WORRALL, D., HOOGEBOOM, E. and WELLING, M. (2019). Learning likelihoods with conditional normalizing flows. arXiv preprint. Available at arXiv:1912.00042.

YANG, Y., YIH, W.-T. and MEEK, C. (2015). Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing* 2013–2018.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1283–1300
https://doi.org/10.1214/26-AOAS2156

EVALUATING A MULTIPLEX DIAGNOSTIC TEST USING PARTIALLY ORDERED BAYES CLASSIFIER

BY YING KUEN CHEUNG[1,a] AND LOUISE KUHN[2,b]

[1]*Department of Biostatistics, Columbia University,* [a]*yc632@cumc.columbia.edu*
[2]*Department of Epidemiology, Columbia University,* [b]*lk24@cumc.columbia.edu*

In a diagnostic test using multiplex assay, each individual biomarker is often expected to have monotonic association with the disease outcome, and, therefore, the underlying disease classification rule is partially ordered with respect to the biomarkers. Nonparametric estimation of the classification rule can be accomplished by projecting an unconstrained Bayes estimator onto the partial ordering subspace. However, computing the projection is challenging as it involves performing maximization over a constrained parameter space whose size grows exponentially with the sample size. We introduce a novel sequential update method for projection-based nonparametric estimation of the disease classification rule and propose new recursive algorithms to implement the method. The proposed algorithms yields the exact Bayes solution that maximizes the posterior gain with respect to a classification-type gain function. When compared to an existing algorithm that gives approximate Bayes solution, our algorithms accomplish the same tasks with much reduced elapsed time in simulation study. We apply the sequential update method to evaluate a human papillomavirus test for cervical cancer precursor lesions and derive diagnostic rule that improves accuracy on existing estimation methods including parametric logistic regression and monotone generalized additive models.

REFERENCES

AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26** 641–647. MR0073895 https://doi.org/10.1214/aoms/1177728423

BACCHETTI, P. (1989). Additive isotonic models. *J. Amer. Statist. Assoc.* **84** 289–294. MR0999691

BISHOP, C. M. and BISHOP, H. (2024). *Deep Learning—Foundations and Concepts*. Springer, Cham. MR4719738 https://doi.org/10.1007/978-3-031-45468-4

BORNKAMP, B., ICKSTADT, K. and DUNSON, D. (2010). Stochastically ordered multiple regression. *Biostatistics* **11** 419–431.

BRUNK, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.* **26** 607–616. MR0073894 https://doi.org/10.1214/aoms/1177728420

CHEUNG, Y. K., CHANDERENG, T. and DIAZ, K. M. (2022). A novel framework to estimate multidimensional minimum effective doses using asymmetric posterior gain and ϵ-tapering. *Ann. Appl. Stat.* **16** 1445–1458. MR4455888 https://doi.org/10.1214/21-aoas1549

CHEUNG, Y. K. and DIAZ, K. M. (2023). Monotone response surface of multi-factor condition: Estimation and Bayes classifiers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 497–522. MR4726975 https://doi.org/10.1093/jrsssb/qkad014

CHEUNG, Y. K and KUHN, L. (2026). Supplement to "Evaluating a multiplex diagnostic test using partially ordered Bayes classifier." https://doi.org/10.1214/26-AOAS2156SUPP

DELONG, E. R., DELONG, D. M. and CLARK-PEARSON, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44** 837–845.

DIAMOND, S. and BOYD, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *J. Mach. Learn. Res.* **17** 83. MR3517106

FU, A., NARASIMHAN, B. and BOYD, S. (2020). CVXR: An R package for disciplined convex optimization. *J. Stat. Softw.* **94** 1–34.

Key words and phrases. Bayes classifier, HPV, multiplex assay, projection, recursive algorithm, sequential update.

JANG, J.-Y., OH, H.-S., LIM, Y. and CHEUNG, Y. K. (2021). Ensemble clustering for step data via binning. *Biometrics* **77** 293–304. MR4229740 https://doi.org/10.1111/biom.13258

JIN, H. and LU, Y. (2009). The optimal linear combination of multiple predictors under the generalized linear models. *Statist. Probab. Lett.* **79** 2321–2327. MR2556363 https://doi.org/10.1016/j.spl.2009.08.002

KUHN, L., SAIDU, R., BOA, R., TERGAS, A., MOODLEY, J., PERSING, D., CAMPBELL, S., TSAI, W. Y. and WRIGHT, T. C. (2020). Clinical evaluation of modifications to a human papillomavirus assay to optimise its utility for cervical cancer screening in low-resource settings: A diagnostic accuracy study. *Lancet Glob. Health* **8** e296–304.

LIN, L. and DUNSON, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika* **101** 303–317. MR3215349 https://doi.org/10.1093/biomet/ast063

LIU, C., LIU, A. and HALABI, S. (2011). A min-max combination of biomarkers to improve diagnostic accuracy. *Stat. Med.* **30** 2005–2014. MR2829061 https://doi.org/10.1002/sim.4238

MANDER, A. P. and SWEETING, M. J. (2015). A product of independent beta probabilities dose escalation design for dual-agent phase I trials. *Stat. Med.* **34** 1261–1276. MR3322767 https://doi.org/10.1002/sim.6434

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

PEPE, M. S. and THOMPSON, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1** 123–140.

RAMSAY, J. O. (1988). Monotone regression splines in action. *Statist. Sci.* **3** 425–441.

RICHARDS, R. J., HAMMITT, J. K. and TSEVAT, J. (1996). Finding the optimal multiple-test strategy using a method analogous to logistic regression. *Med. Decis. Mak.* **16** 367–375.

SERVÉN, D. and BRUMMITT, C. (2018). pyGAM: Generalized Additive Models in Python.

SU, J. Q. and LIU, J. S. (1993). Linear combinations of multiple diagnostic markers. *J. Amer. Statist. Assoc.* **88** 1350–1355. MR1245369

WAN, M. and MCAULEY, J. (2018). Item recommendation on monotonic behavior chain. In *Proceedings of the* 12*th ACM Conference on Recommender Systems* 86–94.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1301–1318
https://doi.org/10.1214/26-AOAS2182

HIERARCHICAL MODELING OF LONGITUDINAL BIOMARKER DATA WITH CHANGEPOINT AND FLEXIBLE SIGMOIDAL RESPONSE

BY MICHELLE NORRIS[1,a], EDWARD J. BEDRICK[2,b], IAN GARDNER[3,c] AND WESLEY JOHNSON[4,d]

[1]*Department of Mathematics and Statistics, Sacramento State University,* [a]*norris@csus.edu*
[2]*Department of Epidemiology and Biostatistics, University of Arizona,* [b]*edwardjbedrick@arizona.edu*
[3]*Atlantic Veterinary College, University of Prince Edward Island,* [c]*iagardner@upei.ca*
[4]*Department of Statistics, University of California, Irvine,* [d]*wjohnson@uci.edu*

We develop a Bayesian hierarchical model for bivariate longitudinal diagnostic outcome data involving testing for the infective agent for Johne's disease (JD). We consider the situation where an imperfect binary test (fecal culture, FC) is repeatedly administered to each individual together with a continuous biomarker (serum ELISA measured as optical density (OD)). For infected individuals we assume the existence of a change-point corresponding to time of infection and posit appropriate changes to model the subsequent responses. Our data consist of records from 12 dairy herds known to be infected with JD. Data were collected from 1984–2003, and tests were performed about every six months. Our joint model incorporates latent states for uninfected and infected cows. We model the serology scores in the infected class using a four-parameter sigmoidal function, with parameters for the unknown time to infection, lag time (time for infective response), and the horizontal asymptote and rate of change. Parametric prior distributions are considered for all unknowns, except the asymptote and rate of change of the sigmoidal curves which are modeled with Dirichlet process mixtures and which allows for clustering of shapes of serologic response curves.

REFERENCES

ALSEFRI, M., SUDELL, M., GARCÍA-FIÑANA, M. and KOLAMUNNAGE-DONA, R. (2020). Bayesian joint modelling of longitudinal and time to event data: A methodological review. *BMC Med. Res. Methodol.* **20** 1–17.

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

BRANSCUM, A., GARDNER, I. and JOHNSON, W. (2004). Bayesian modeling of animal-and herd-level prevalences. *Prev. Vet. Med.* **66** 101–112.

BRANSCUM, A. J., GARDNER, I. A. and JOHNSON, W. O. (2005). Estimation of diagnostic test sensitivity and specificity through Bayesian modeling. *Prev. Vet. Med.* **68** 145–163.

ČERNILA, M., LOGAR, M., MOŽINA, H. and OSREDKAR, J. (2023). Comparison between the sofia SARS antigen FIA test and the PCR test in detection of SARS-CoV-2 infection. *Lab. Med.* **54** e44–e48.

CHOI, Y. K., JOHNSON, W. O., COLLINS, M. T. and GARDNER, I. A. (2006). Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. *J. Agric. Biol. Environ. Stat.* **11** 210–229.

CHRISTENSEN, R., JOHNSON, W., BRANSCUM, A. and HANSON, T. E. (2010). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians. Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR2682928

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

ENØE, C., GEORGIADIS, M. P. and JOHNSON, W. O. (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev. Vet. Med.* **45** 61–81.

GILES, J., JOHNSON, W., JONES, G., HEUER, C. and DUNOWSKA, M. (2018). Development of an indirect ELISA for detection of antibody to wobbly possum disease virus in archival sera of Australian brushtail possums (Trichosurus vulpecula) in New Zealand. *N. Z. Vet. J.* **66** 186–193.

Key words and phrases. Bayes' Theorem, Dirichlet process mixture, reversible jump MCMC, Johne's disease, sigmoidal serologic response, sensitivity, specificity, knowledge based priors.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810 https://doi.org/10.1093/biomet/82.4.711

HANSON, T., JOHNSON, W. O. and GARDNER, I. A. (2003). Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold standard. *J. Agric. Biol. Environ. Stat.* **8** 223–239.

HANSON, T. E., BRANSCUM, A. J. and JOHNSON, W. O. (2011). Predictive comparison of joint longitudinal-survival modeling: A case study illustrating competing approaches. *Lifetime Data Anal.* **17** 3–28. Supplementary material available online. MR2764577 https://doi.org/10.1007/s10985-010-9162-0

HORROCKS, J. and VAN DEN HEUVEL, M. J. (2009). Prediction of pregnancy: A joint model for longitudinal and binary data. *Bayesian Anal.* **4** 523–538. MR2551044 https://doi.org/10.1214/09-BA419

JOHNSON, W. O., JONES, G. and GARDNER, I. A. (2019). Gold standards are out and Bayes is in: Implementing the cure for imperfect reference tests in diagnostic accuracy studies. *Prev. Vet. Med.* **167** 113–127.

JONES, G., JOHNSON, W. O., VINK, W. D. and FRENCH, N. (2012). A framework for the joint modeling of longitudinal diagnostic outcome data and latent infection status: Application to investigating the temporal relationship between infection and disease. *Biometrics* **68** 371–379. MR2959603 https://doi.org/10.1111/j.1541-0420.2011.01687.x

LEPPER, A., WILKS, C., KOTIW, M., WHITEHEAD, J. and SWART, K. (1989). Sequential bacteriological observations in relation to cell-mediated and humoral antibody responses of cattle infected with Mycobacterium paratuberculosis and maintained on normal or high iron intake. *Aust. Vet. J.* **66** 50–55.

LI, E., WANG, N. and WANG, N.-Y. (2007). Joint models for a primary endpoint and multiple longitudinal covariate processes. *Biometrics* **63** 1068–1078. MR2414584 https://doi.org/10.1111/j.1541-0420.2007.00822.x

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

NORRIS, M., BEDRICK, E. J., GARDNER, I. and JOHNSON, W. (2026). Supplement to "Hierarchical modeling of longitudinal biomarker data with changepoint and flexible sigmoidal response." https://doi.org/10.1214/26-AOAS2182SUPPA, https://doi.org/10.1214/26-AOAS2182SUPPB

NORRIS, M., JOHNSON, W. O. and GARDNER, I. A. (2009). Modeling bivariate longitudinal diagnostic outcome data in the absence of a gold standard. *Stat. Interface* **2** 171–185. MR2516068 https://doi.org/10.4310/SII.2009.v2.n2.a7

NORRIS, M., JOHNSON, W. O. and GARDNER, I. A. (2014). Bayesian semi-parametric joint modeling of biomarker data with a latent changepoint: Assessing the temporal performance of enzyme-linked immunosorbent assay (ELISA) testing for paratuberculosis. *Stat. Interface* **7** 417–438. MR3302372 https://doi.org/10.4310/SII.2014.v7.n4.a1

QUINTANA, F. A., JOHNSON, W. O., WAETJEN, L. E. and GOLD, E. B. (2016). Bayesian nonparametric longitudinal data analysis. *J. Amer. Statist. Assoc.* **111** 1168–1181. MR3561940 https://doi.org/10.1080/01621459.2015.1076725

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

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

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1319–1339
https://doi.org/10.1214/26-AOAS2181

A BAYESIAN APPROACH FOR SELECTING RELEVANT EXTERNAL DATA (BASE): APPLICATION TO A STUDY OF LONG-TERM OUTCOMES IN A HEMOPHILIA GENE THERAPY TRIAL (HOPE-B)

BY TIANYU PAN[1,a], YIYAO SHI[2,b], XIANG ZHANG[3,d], WEINING SHEN[2,c] AND TING YE[4,e]

[1]*Department of Biomedical Data Science, Stanford University,* [a]*tianyup2@stanford.edu*
[2]*Department of Statistics, University of California, Irvine,* [b]*yiyaos@uci.edu,* [c]*weinings@uci.edu*
[3]*Medical Affairs and HTA Statistics, Quantitative Clinical Sciences and Reporting, CSL Behring,* [d]*Xiang.Zhang@cslbehring.com*
[4]*Department of Biostatistics, University of Washington,* [e]*tingye1@uw.edu*

Gene therapies aim to address the root causes of diseases, particularly those stemming from rare genetic defects that can be life-threatening or severely debilitating. Although an increasing number of gene therapies have received regulatory approvals in recent years, understanding their long-term efficacy in trials with limited follow-up time remains challenging. To address this question, we propose a novel Bayesian framework to selectively integrate relevant external data with internal trial data to improve the inference of the durability of long-term efficacy. We proved that the proposed method can theoretically identify external subsets deemed relevant, where relevance is defined as the similarity, induced by the marginal likelihood, between the generating mechanisms of the internal data and the selected external data. We conducted simulations to evaluate its performance under various scenarios. Furthermore, we apply this method to predict and infer the endogenous factor IX (FIX) levels of patients who receive Etranacogene dezaparvovec long term. Our estimated long-term FIX levels, validated by recent trial data, indicate that Etranacogene dezaparvovec induces sustained FIX production. Together, the theoretical findings, simulation results, and application of this framework underscore its potential to address long-term effectiveness estimation and inference questions in real-world applications.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification: System identification and time-series analysis. *IEEE Trans. Automat. Control* **AC-19** 716–723. MR0423716 https://doi.org/10.1109/tac.1974.1100705

ALT, E. M., CHANG, X., JIANG, X., LIU, Q., MO, M., XIA, H. A. and IBRAHIM, J. G. (2024). LEAP: The latent exchangeability prior for borrowing information from historical data. *Biometrics* **80** Paper No. ujae083. MR4896882 https://doi.org/10.1093/biomtc/ujae083

ATHEY, S., CHETTY, R., IMBENS, G. W. and KANG, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely Technical Report National Bureau of Economic Research.

BLANCHETTE, V., KEY, N., LJUNG, L., MANCO-JOHNSON, M., VAN DEN BERG, H. and SRIVASTAVA, A. (2014). Definitions in hemophilia: Communication from the SSC of the ISTH. *J. Thromb. Haemost.* **12** 1935–1939.

CAI, T., CAI, T. T. and ZHANG, A. (2016). Structured matrix completion with applications to genomic data integration. *J. Amer. Statist. Assoc.* **111** 621–633. MR3538692 https://doi.org/10.1080/01621459.2015.1021005

CHEN, M.-H., IBRAHIM, J. G. and SHAO, Q.-M. (2000). Power prior distributions for generalized linear models. *J. Statist. Plann. Inference* **84** 121–137. MR1747500 https://doi.org/10.1016/S0378-3758(99)00140-8

CHEN, M.-H., IBRAHIM, J. G. and YIANNOUTSOS, C. (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 223–242. MR1664057 https://doi.org/10.1111/1467-9868.00173

Key words and phrases. Bayesian analysis, data integration, gene therapy, long-term outcome inference, selective borrowing.

CHEN, S., LI, S., ZHANG, B. and YE, T. (2025). Minimax rates and adaptivity in combining experimental and observational data. *J. Causal Inference* **13** Paper No. 20240024. MR4933607 https://doi.org/10.1515/jci-2024-0024

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

CHU, Y. and YUAN, Y. (2018). A Bayesian basket trial design using a calibrated Bayesian hierarchical model. *Clin. Trials* **15** 149–158.

DAI, X. and LI, L. (2023). Orthogonalized kernel debiased machine learning for multimodal data analysis. *J. Amer. Statist. Assoc.* **118** 1796–1810. MR4646607 https://doi.org/10.1080/01621459.2021.2013851

KEE, A. and MAIO, V. (2019). Value-based contracting: Challenges and opportunities. *Am. J. Med. Qual.* **34** 615–617.

FARMER, C., NIKRAM, E., TRIGG, L. A., ROBINSON, S., COPPELL, J., MUTHUKUMAR, M., MELENDEZ-TORRES, G. J. and WILSON, E. C. (2023). Etranacogene dezaparvovec for treating moderately severe or severe haemophilia B. *Nat. Inst. Health Care Excell.*.

FRAZIER, D. T., NOTT, D. J. and DROVANDI, C. (2025). Synthetic likelihood in misspecified models. *J. Amer. Statist. Assoc.* **120** 884–895. MR4917264 https://doi.org/10.1080/01621459.2024.2370594

GEORGE, L. A., SULLIVAN, S. K., GIERMASZ, A., RASKO, J. E., SAMELSON-JONES, B. J., DUCORE, J., CUKER, A., SULLIVAN, L. M., MAJUMDAR, S. et al. (2017). Hemophilia B gene therapy with a high-specific-activity factor IX variant. *N. Engl. J. Med.* **377** 2215–2227.

GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834

GOMEZ, E., GIERMASZ, A., CASTAMAN, G., KEY, N. S., LATTIMORE, S. U., LEEBEEK, F. W., MIESBACH, W., RECHT, M., DRYGALSKI, A. et al. (2021). Etranacogene Dezaparvovec (AAV5-Padua hFIX variant, AMT-061), an enhanced vector for gene transfer in adults with severe or moderate-severe hemophilia B: 2.5 year data from a phase 2b trial. *ISTH Congress*.

HASSLER, G. W., MAGEE, A. F., ZHANG, Z., BAELE, G., LEMEY, P., JI, X., FOURMENT, M. and SUCHARD, M. A. (2023). Data integration in Bayesian phylogenetics. *Annu. Rev. Stat. Appl.* **10** 353–377. MR4567797 https://doi.org/10.1146/annurev-statistics-033021-112532

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

HOBBS, B. P. and LANDIN, R. (2018). Bayesian basket trial design with exchangeability monitoring. *Stat. Med.* **37** 3557–3572. MR3869118 https://doi.org/10.1002/sim.7893

IBRAHIM, J. G., CHEN, M.-H., GWON, Y. and CHEN, F. (2015). The power prior: Theory and applications. *Stat. Med.* **34** 3724–3749. MR3422144 https://doi.org/10.1002/sim.6728

IMBENS, G., KALLUS, N., MAO, X. and WANG, Y. (2025). Long-term causal inference under persistent confounding via data combination. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **87** 362–388. MR4896650 https://doi.org/10.1093/jrsssb/qkae095

LEEBEEK, F. W., MEIJER, K., COPPENS, M., KAMPMANN, P., VAN DER VALK, P., MONAHAN, P. E., PINACHYAN, K., LEQUELLEC, S., MIESBACH, W. et al. (2024). Stable factor IX expression and sustained reductions in factor IX use 8 years after gene therapy with CSL220 (formerly AMT-060) in adults with hemophilia B. *Blood* **144** 3578.

LEEBEEK, F. W. and MIESBACH, W. (2021). Gene therapy for hemophilia: A review on clinical benefit, limitations, and remaining issues. *Blood* **138** 923–931.

LI, Q. and LI, L. (2022). Integrative factor regression and its inference for multimodal data analysis. *J. Amer. Statist. Assoc.* **117** 2207–2221. MR4528499 https://doi.org/10.1080/01621459.2021.1914635

LI, S. and LUEDTKE, A. (2023). Efficient estimation under data fusion. *Biometrika* **110** 1041–1054. MR4667439 https://doi.org/10.1093/biomet/asad007

LIU, Y., SUN, X., ZHONG, W. and LI, B. (2022). B-scaling: A novel nonparametric data fusion method. *Ann. Appl. Stat.* **16** 1292–1312. MR4455881 https://doi.org/10.1214/21-aoas1537

MIESBACH, W., MEIJER, K., COPPENS, M., KAMPMANN, P., KLAMROTH, R., SCHUTGENS, R., CASTAMAN, G., SAWYER, E. and LEEBEEK, F. (2021). Five year data confirms stable FIX expression and sustained reductions in bleeding and factor IX use following AMT-060 gene therapy in adults with severe or moderate-severe hemophilia B. *ISTH Congress*.

NATHWANI, A. C., REISS, U., TUDDENHAM, E., CHOWDARY, P., MCINTOSH, J., RIDDELL, A., PIE, J., MAHLANGU, J. N., RECHT, M. et al. (2018). Adeno-associated mediated gene transfer for hemophilia B:8 year follow up and impact of removing "empty viral particles" on safety and efficacy of gene transfer. *Blood* **132** 491–491.

NATHWANI, A. C., REISS, U. M., TUDDENHAM, E. G., ROSALES, C., CHOWDARY, P., MCINTOSH, J., DELLA PERUTA, M., LHERITEAU, E., PATEL, N. et al. (2014). Long-term safety and efficacy of factor IX gene therapy in hemophilia B. *N. Engl. J. Med.* **371** 1994–2004.

NATHWANI, A. C., TUDDENHAM, E. G., RANGARAJAN, S., ROSALES, C., MCINTOSH, J., LINCH, D. C., CHOWDARY, P., RIDDELL, A., PIE, A. J. et al. (2011). Adenovirus-associated virus vector–mediated gene transfer in hemophilia B. *N. Engl. J. Med.* **365** 2357–2365.

NEUENSCHWANDER, B., BRANSON, M. and SPIEGELHALTER, D. J. (2009). A note on the power prior. *Stat. Med.* **28** 3562–3566. MR2744383 https://doi.org/10.1002/sim.3722

NEUENSCHWANDER, B., WANDEL, S., ROYCHOUDHURY, S. and BAILEY, S. (2016). Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharm. Stat.* **15** 123–134.

NEWTON, M. A. and RAFTERY, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. Roy. Statist. Soc. Ser. B, Methodol.* **56** 3–48. MR1257793

OBERST, M., D'AMOUR, A., CHEN, M., WANG, Y., SONTAG, D. and YADLOWSKY, S. (2022). Bias-robust integration of observational and experimental estimators. arXiv Preprint. Available at arXiv:2205.10467.

PAJOR, A. (2017). Estimating the marginal likelihood using the arithmetic mean identity. *Bayesian Anal.* **12** 261–287. MR3620130 https://doi.org/10.1214/16-BA1001

PAN, T., SHI, Y., ZHANG, X., SHEN, W. and YE, T. (2026). Supplement to "A Bayesian approach for selecting relevant external data (BASE): application to a study of long-term outcomes in a hemophilia gene therapy trial (HOPE-B)." https://doi.org/10.1214/26-AOAS2181SUPP

PARK, S., XU, H. and ZHAO, H. (2021). Integrating multidimensional data for clustering analysis with applications to cancer patient data. *J. Amer. Statist. Assoc.* **116** 14–26. MR4227671 https://doi.org/10.1080/01621459.2020.1730853

PIPE, S. W., LEEBEEK, F. W., RECHT, M., KEY, N. S., CASTAMAN, G., MIESBACH, W., LATTIMORE, S., PEERLINCK, K., VAN DER VALK, P. et al. (2023). Gene therapy with etranacogene dezaparvovec for hemophilia B. *N. Engl. J. Med.* **388** 706–718.

PIPE, S. W., MIESBACH, W., RECHT, M., LEEBEEK, F. W., KEY, N. S., CASTAMAN, G., LATTIMORE, S., COPPENS, M., LE QUELLEC, S. et al. (2025). Final analysis of a study of etranacogene dezaparvovec for hemophilia B. *N. Engl. J. Med.*.

PSIODA, M. A., XU, J., JIANG, Q., KE, C., YANG, Z. and IBRAHIM, J. G. (2021). Bayesian adaptive basket trial design using model averaging. *Biostatistics* **22** 19–34. MR4207142 https://doi.org/10.1093/biostatistics/kxz014

REISS, U. M., DAVIDOFF, A. M., TUDDENHAM, E. G., CHOWDARY, P., MCINTOSH, J., MUCZYNSKI, V., JOURNOU, M., SIMINI, G., IRELAND, L. et al. (2025). Sustained clinical benefit of AAV gene therapy in severe hemophilia B. *N. Engl. J. Med.* **392** 2226–2234.

RITCHIE, M. D., HOLZINGER, E. R., LI, R., PENDERGRASS, S. A. and KIM, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* **16** 85–97.

ROBERT, C. P. (2007). *The Bayesian Choice*: *From Decision-Theoretic Foundations to Computational Implementation*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR2723361

SAMELSON-JONES, B. J., SULLIVAN, S. K., RASKO, J. E., GIERMASZ, A., GEORGE, L. A., DUCORE, J. M., TEITEL, J. M., MCGUINN, C. E., O'BRIEN, A. et al. (2021). Follow-up of more than 5 years in a cohort of patients with hemophilia B treated with fidanacogene elaparvovec adeno-associated virus gene therapy. *Blood* **138** 3975.

SCHMIDLI, H., GSTEIGER, S., ROYCHOUDHURY, S., O'HAGAN, A., SPIEGELHALTER, D. and NEUENSCHWANDER, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* **70** 1023–1032. MR3295763 https://doi.org/10.1111/biom.12242

SHAH, J., KIM, H., SIVAMURTHY, K., MONAHAN, P. E. and FRIES, M. (2023). Comprehensive analysis and prediction of long-term durability of factor IX activity following etranacogene dezaparvovec gene therapy in the treatment of hemophilia B. *Curr. Med. Res. Opin.* **39** 227–237.

SHEN, W., TOKDAR, S. T. and GHOSAL, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. MR3094441 https://doi.org/10.1093/biomet/ast015

SRIVASTAVA, A., SANTAGOSTINO, E., DOUGALL, A., KITCHEN, S., SUTHERLAND, M., PIPE, S. W., CARCAO, M., MAHLANGU, J., RAGNI, M. V. et al. (2020). WFH guidelines for the management of hemophilia. *Haemophilia* **26** 1–158.

TICE, J., WALTON, S., HERCE-HAGIWARA, B., FAHIM, S., MORADI, A., SARKER, J., CHU, J., AGBOOLA, F., PEARSON, S. et al. (2022). Gene therapy for hemophilia B and an update on gene therapy for hemophilia A: Effectiveness and value. *Inst. Clin. Econ. Rev.*.

XUE, F. and QU, A. (2021). Integrating multisource block-wise missing data in model selection. *J. Amer. Statist. Assoc.* **116** 1914–1927. MR4353722 https://doi.org/10.1080/01621459.2020.1751176

YANG, S., GAO, C., ZENG, D. and WANG, X. (2023). Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 575–596. MR4728022 https://doi.org/10.1093/jrsssb/qkad017

YANG, S., KIM, J. K. and SONG, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 445–465. MR4084171

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1340–1363
https://doi.org/10.1214/26-AOAS2186

INFERENCE ON SUMMARIES OF A MODEL-AGNOSTIC LONGITUDINAL VARIABLE IMPORTANCE TRAJECTORY WITH APPLICATION TO SUICIDE PREVENTION

BY BRIAN D WILLIAMSON[1,2,3,a], ERICA E. M. MOODIE[4,c], GREGORY E. SIMON[5,6,7,d], REBECCA C. ROSSOM[8,e] AND SUSAN M. SHORTREED[1,3,b]

[1]*Biostatistics Division, Kaiser Permanente Washington Health Research Institute,* [a]*brian.d.williamson@kp.org,* [b]*susan.m.shortreed@kp.org*

[2]*Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center*

[3]*Department of Biostatistics, University of Washington*

[4]*Department of Epidemiology, Biostatistics, and Occupational Health, McGill University,* [c]*erica.moodie@mcgill.ca*

[5]*Investigative Science Division, Kaiser Permanente Washington Health Research Institute*

[6]*Department of Health Systems Science, Kaiser Permanente Bernard J. Tyson School of Medicine*

[7]*Department of Psychiatry and Behavioral Sciences, University of Washington,* [d]*gregory.e.simon@kp.org*

[8]*Division of Research, HealthPartners Institute,* [e]*rebecca.c.rossom@healthpartners.com*

Risk of suicide attempt varies over time. Understanding the importance of risk factors measured at a mental health visit can help clinicians evaluate future risk and provide appropriate care during the visit. In prediction settings where data are collected over time, such as in mental health care, it is often of interest to understand both the importance of variables for predicting the response at each time point and the importance summarized over the time series. Building on recent advances in estimation and inference for variable importance measures, we define summaries of variable importance trajectories and corresponding estimators. Under common regularity conditions, the same approaches for inference can be applied to these measures regardless of the choice of the algorithm(s) used to estimate the prediction function under standard convergence conditions. We propose a nonparametric efficient estimation and inference procedure as well as a null hypothesis testing procedure that are valid even when complex machine learning tools are used for prediction. Through simulations we demonstrate that our proposed procedures have good operating characteristics. We use these approaches to analyze electronic health records data from two large health systems to investigate the longitudinal importance of risk factors (i.e., the importance of risk factors over time) for suicide attempt to inform future suicide prevention research and clinical workflow.

REFERENCES

ADAMS, R., HAROZ, E. E., REBMAN, P., SUTTLE, R., GROSVENOR, L., BAJAJ, M., DAYAL, R. R., MAGGIO, D., KETTERING, C. L. et al. (2024). Developing a suicide risk model for use in the Indian Health Service. *npj Mental Health Res*. **3** 47.

BARAK-CORREN, Y., CASTRO, V. M., JAVITT, S., HOFFNAGLE, A. G., DAI, Y., PERLIS, R. H., NOCK, M. K., SMOLLER, J. W. and REIS, B. Y. (2017). Predicting suicidal behavior from longitudinal electronic health records. *Amer. J. Psychiatry* **174** 154–162.

BAYRAMLI, I., CASTRO, V., BARAK-CORREN, Y., MADSEN, E. M., NOCK, M. K., SMOLLER, J. W. and REIS, B. Y. (2022). Temporally informed random forests for suicide risk prediction. *J. Amer. Med. Inform. Assoc*. **29** 62–71.

BIAN, Z., MOODIE, E. E. M., SHORTREED, S. M. and BHATNAGAR, S. (2023). Variable selection in regression-based estimation of dynamic treatment regimes. *Biometrics* **79** 988–999. MR4606331 https://doi.org/10.1111/biom.13608

Key words and phrases. Intrinsic variable importance, longitudinal data, machine learning, prediction, risk factors of self-harm, suicide prevention.

BOILEAU, P., QI, N. T., VAN DER LAAN, M. J., DUDOIT, S. and LENG, N. (2023). A flexible approach for predictive biomarker discovery. *Biostatistics* **24** 1085–1105. MR4656777 https://doi.org/10.1093/biostatistics/kxac029

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

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

CHEN, Q., ZHANG-JAMES, Y., BARNETT, E. J., LICHTENSTEIN, P., JOKINEN, J., D'ONOFRIO, B. M., FARAONE, S. V., LARSSON, H. and FAZEL, S. (2020). Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: A machine learning study using Swedish national registry data. *PLoS Med.* **17** e1003416.

CHEN, S., TIAN, L., CAI, T. and YU, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* **73** 1199–1209. MR3744534 https://doi.org/10.1111/biom.12676

COLEY, R. Y., JOHNSON, E., SIMON, G. E., CRUZ, M. and SHORTREED, S. M. (2021). Racial/ethnic disparities in the performance of prediction models for death by suicide after mental health visits. *JAMA Psychiatr.* **78** 726–734.

COLEY, R. Y., LIAO, Q., SIMON, N. and SHORTREED, S. M. (2023). Empirical evaluation of internal validation methods for prediction in large-scale clinical data with rare-event outcomes: A case study in suicide risk prediction. *BMC Med. Res. Methodol.* **23** 33.

COULOMBE, J., MOODIE, E. E. M., SHORTREED, S. M. and RENOUX, C. (2023). Estimating individualized treatment rules in longitudinal studies with covariate-driven observation times. *Stat. Methods Med. Res.* **32** 868–884. MR4597912 https://doi.org/10.1177/09622802231158733

DÍAZ, I., HUBBARD, A., DECKER, A. and COHEN, M. (2015). Variable importance and prediction methods for longitudinal problems with missing variables. *PLoS ONE* **10** e0120031.

FAN, A., LU, W. and SONG, R. (2016). Sequential advantage selection for optimal treatment regime. *Ann. Appl. Stat.* **10** 32–53. MR3480486 https://doi.org/10.1214/15-AOAS849

FIELD, C. A. and WELSH, A. H. (2007). Bootstrapping clustered data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 369–390. MR2323758 https://doi.org/10.1111/j.1467-9868.2007.00593.x

FISHER, A., RUDIN, C. and DOMINICI, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* **20** Paper No. 177, 81. MR4048988

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451

GRADUS, J. L., ROSELLINI, A. J., HORVÁTH-PUHÓ, E., STREET, A. E., GALATZER-LEVY, I., JIANG, T., LASH, T. L. and SØRENSEN, H. T. (2020). Prediction of sex-specific suicide risk using machine learning and single-payer health care registry data from Denmark. *JAMA Psychiatr.* **77** 25–34.

GRÖMPING, U. (2015). Variable importance in regression models. *Wiley Interdiscip. Rev.: Comput. Stat.* **7** 137–152. MR3349293 https://doi.org/10.1002/wics.1346

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

HAROZ, E. E., GRUBIN, F., GOKLISH, N., PIOCHE, S., CWIK, M., BARLOW, A., WAUGH, E., USHER, J., LENERT, M. C. et al. (2021). Designing a clinical decision support tool that leverages machine learning for suicide risk prediction: Development study in partnership with Native American care providers. *JMIR Public Health Surveill.* **7** e24377.

HAROZ, E. E., REBMAN, P., GOKLISH, N., GARCIA, M., SUTTLE, R., MAGGIO, D., CLATTENBURG, E., MEGA, J. and ADAMS, R. (2024). Performance of machine learning suicide risk models in an American Indian population. *JAMA Netw. Open* **7** e2439269–e2439269.

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

HINES, O. J., DIAZ-ORDAZ, K. and VANSTEELANDT, S. (2025). Variable importance measures for heterogeneous treatment effects. *Biometrics* **81** Paper No. ujaf140. MR5006513 https://doi.org/10.1093/biomtc/ujaf140

HOROWITZ, L. M., BRIDGE, J. A., TEACH, S. J., BALLARD, E., KLIMA, J., ROSENSTEIN, D. L., WHARFF, E. A., GINNIS, K., CANNON, E. et al. (2012). Ask Suicide-Screening Questions (ASQ): A brief instrument for the pediatric emergency department. *JAMA Pediatr.* **166** 1170–1176.

HUBBARD, A., DÍAZ MUÑOZ, I., DECKER, A., HOLCOMB, J., SCHREIBER, M., BULGER, E. et al. (2013). Time-dependent prediction and evaluation of variable importance using superlearning in high dimensional clinical data. *J. Trauma Acute Care Surg.* **75** S53–S60.

ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* **2** 841–860. MR2516796 https://doi.org/10.1214/08-AOAS169

JACOBS, D. G., BALDESSARINI, R. J., CONWELL, Y., FAWCETT, J. A., HORTON, L., MELTZER, H. et al. (2010). Assessment and treatment of patients with suicidal behaviors. *APA Pract. Guidel.* **1** 183.

KESSLER, R. C., BOSSARTE, R. M., LUEDTKE, A., ZASLAVSKY, A. M. and ZUBIZARRETA, J. R. (2020). Suicide prediction models: A critical review of recent research with recommendations for the way forward. *Mol. Psychiatry* **25** 168–179.

KESSLER, R. C., STEIN, M. B., PETUKHOVA, M. V., BLIESE, P., BOSSARTE, R. M., BROMET, E. J., FULLERTON, C. S., GILMAN, S. E., IVANY, C. et al. (2017). Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *Mol. Psychiatry* **22** 544–551.

KESSLER, R. C., WARNER, L. C. H., IVANY, L. C., PETUKHOVA, M. V., ROSE, S., BROMET, E. J., BROWN, L. M. III, CAI, T., COLPE, L. J. et al. (2015). Predicting US Army suicides after hospitalizations with psychiatric diagnoses in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *JAMA Psychiatr.* **72** 49.

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

KROENKE, K., SPITZER, R. L., WILLIAMS, J. B. and LÖWE, B. (2010). The patient health questionnaire somatic, anxiety, and depressive symptom scales: A systematic review. *Gen. Hosp. Psychiatry* **32** 345–359.

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

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

LINK, W. and SAUER, J. (1997). Estimation of population trajectories from count data. *Biometrics* **53** 488–497.

LUNDBERG, S. and LEE, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* **30**.

MA, T., CAI, H., QI, Z., SHI, C. and LABER, E. (2023). Sequential knockoffs for variable selection in reinforcement learning. arXiv preprint. Available at arXiv:2303.14281.

MATARAZZO, B. B., EAGAN, A., LANDES, S. J., MINA, L. K. et al. (2023). The Veterans Health Administration REACH VET program: Suicide predictive modeling in practice. *Psychiatr. Serv.* **74** 206–209.

MINUS, E., COLEY, R. Y., SHORTREED, S. M. and WILLIAMSON, B. D. (2025). Behavior of prediction performance metrics with rare events. *J. Clin. Epidemiol.* https://doi.org/10.1016/j.jclinepi.2025.112046

MUNDT, J., GRIEST, J., JEFFERSON, J., FEDERICO, M., MANN, J. and POSNER, K. (2013). Prediction of suicidal behavior in clinical research by lifetime suicidal ideation and behavior ascertained by the electronic Columbia-Suicide Severity Rating Scale. *J. Clin. Psychiatry* **74** 887–893.

MURDOCH, W. J., SINGH, C., KUMBIER, K., ABBASI-ASL, R. and YU, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA* **116** 22071–22080. MR4030584 https://doi.org/10.1073/pnas.1900654116

NOCK, M. K., MILLNER, A. J., ROSS, E. L., KENNEDY, C. J., AL-SUWAIDI, M., BARAK-CORREN, Y., CASTRO, V. M., CASTRO-RAMIREZ, F., LAURICELLA, T. et al. (2022). Prediction of suicide attempts using clinician assessment, patient self-report, and electronic health records. *JAMA Netw. Open* **5** e2144373–e2144373.

PAPINI, S., HSIN, H., KIPNIS, P., LIU, V. X., LU, Y., GIRARD, K., STERLING, S. A. and ITURRALDE, E. M. (2024). Validation of a multivariable model to predict suicide attempt in a mental health intake sample. *JAMA Psychiatr.* **81** 700–707.

PAPINI, S., HSIN, H., KIPNIS, P., LIU, V. X., LU, Y., STERLING, S. A. and ITURRALDE, E. (2023). Performance of a prediction model of suicide attempts across race and ethnicity. *JAMA Psychiatr.* **80** 399–400.

PENFOLD, R. B., JOHNSON, E., SHORTREED, S. M., ZIEBELL, R. A., LYNCH, F. L., CLARKE, G. N., COLEMAN, K. J., WAITZFELDER, B. E., BECK, A. L. et al. (2021). Predicting suicide attempts and suicide deaths among adolescents following outpatient visits. *J. Affective Disorders* **294** 39–47.

POSNER, K., BROWN, G. K., STANLEY, B., BRENT, D. A., YERSHOVA, K. V., OQUENDO, M. A., CURRIER, G. W., MELVIN, G. A., GREENHILL, L. et al. (2011). The Columbia–Suicide Severity Rating Scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *Amer. J. Psychiatry* **168** 1266–1277.

ROBINS, J., ORELLANA, L. and ROTNITZKY, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Stat. Med.* **27** 4678–4721. MR2528576 https://doi.org/10.1002/sim.3301

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

ROSSOM, R. C., RICHARDS, J. E., STERLING, S., AHMEDANI, B., BOGGS, J. M., YARBOROUGH, B. J. H., BECK, A., LLOYD, K., FRANK, C. et al. (2022). Connecting research and practice: Implementation of suicide prevention strategies in learning health care systems. *Psychiatr. Serv.* **73** 219–222.

SANDERSON, M., BULLOCH, A. G., WANG, J., WILLIAMS, K. G., WILLIAMSON, T. and PATTEN, S. B. (2020). Predicting death by suicide following an emergency department visit for parasuicide with administrative health care system data and machine learning. *EClinicalMedicine* **20**.

SANDERSON, M., BULLOCH, A. G., WANG, J., WILLIAMSON, T. and PATTEN, S. B. (2019). Predicting death by suicide using administrative health care system data: Can feedforward neural network models improve upon logistic regression models? *J. Affective Disorders* **257** 741–747.

SHAW, J. L., BEANS, J. A., NOONAN, C., SMITH, J. J., MOSLEY, M., LILLIE, K. M., AVEY, J. P., ZIEBELL, R. and SIMON, G. (2022). Validating a predictive algorithm for suicide risk with Alaska Native populations. *Suicide Life-Threat. Behav.* **52** 696–704.

SHORTREED, S., WALKER, R., JOHNSON, E., WELLMAN, R., CRUZ, M., ZIEBELL, R. et al. (2023). Complex modeling with detailed temporal predictors does not improve health records-based suicide risk prediction. *npj Digit. Med.* **6** 47.

SIMON, G., JOHNSON, E., LAWRENCE, J., ROSSOM, R., AHMEDANI, B., LYNCH, F. et al. (2018). Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *Amer. J. Psychiatry* **175** 951–960.

SIMON, G., RUTTER, C., PETERSON, D., OLIVER, M., WHITESIDE, U., OPERSKALSKI, B. and LUDMAN, E. (2013). Does response on the PHQ-9 depression questionnaire predict subsequent suicide attempt or suicide death? *Psychiatr. Serv.* **64** 1195–1202.

SIMON, G., YARBOROUGH, B., ROSSON, R., LAWRENCE, J., LYNCH, F., WAITZFELDER, B., AHMEDANI, B. and SHORTREED, S. (2019a). Self-reported suicidal ideation predicts suicidal behavior in out-patient receiving psychotic disorder diagnosis. *Psychiatr. Serv.* **70** 176–83.

SIMON, G. E., COLEMAN, K. J., ROSSOM, R. C., BECK, A., OLIVER, M., JOHNSON, E., WHITESIDE, U., OPERSKALSKI, B., PENFOLD, R. B. et al. (2016). Risk of suicide attempt and suicide death following completion of the Patient Health Questionnaire depression module in community practice. *J. Clin. Psychiatry* **77** 20461.

SIMON, G. E., CRUZ, M., SHORTREED, S. M., STERLING, S. A., COLEMAN, K. J., AHMEDANI, B. K., YASEEN, Z. S. and MOSHOLDER, A. D. (2024). Stability of suicide risk prediction models during changes in health care delivery. *Psychiatr. Serv.* **75** 139–147.

SIMON, G. E., SHORTREED, S. M., JOHNSON, E., BECK, A., COLEMAN, K. J., ROSSOM, R. C., WHITESIDE, U. S., OPERSKALSKI, B. H. and PENFOLD, R. B. (2017). Between-visit changes in suicidal ideation and risk of subsequent suicide attempt. *Depress. Anxiety* **34** 794–800.

SIMON, G. E., SHORTREED, S. M., JOHNSON, E., ROSSOM, R. C., LYNCH, F. L., ZIEBELL, R. and PENFOLD, R. B. (2019b). What health records data are required for accurate prediction of suicidal behavior? *J. Amer. Med. Inform. Assoc.* **26** 1458–1465.

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

TSUI, F. R., SHI, L., RUIZ, V., RYAN, N. D., BIERNESSER, C., IYENGAR, S., WALSH, C. G. and BRENT, D. A. (2021). Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. *JAMIA Open* **4** ooab011.

VAN DER LAAN, M. J. (2006). Statistical inference for variable importance. *Int. J. Biostat.* **2** Art. 2, 33. MR2275897 https://doi.org/10.2202/1557-4679.1008

VAN DER LAAN, M. J. and LUEDTKE, A. R. (2015). Targeted learning of the mean outcome under an optimal dynamic treatment rule. *J. Causal Inference* **3** 61–95. MR4289428 https://doi.org/10.1515/jci-2013-0022

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

VOGT, D., ROSELLINI, A. J., BOROWSKI, S., STREET, A. E., O'BRIEN, R. W. and TOMOYASU, N. (2024). How well can US military veterans' suicidal ideation be predicted from static and change-based indicators of their psychosocial well-being as they adapt to civilian life? *Soc. Psychiatry Psychiatr. Epidemiol.* **59** 261–271.

WALKER, R. L., SHORTREED, S. M., ZIEBELL, R. A., JOHNSON, E., BOGGS, J. M., LYNCH, F. L., DAIDA, Y. G., AHMEDANI, B. K., ROSSOM, R. et al. (2021). Evaluation of electronic health record-based suicide risk prediction models on contemporary data. *Appl. Clin. Inform.* **12** 778–787.

WALSH, C. G., RIBEIRO, J. D. and FRANKLIN, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clin. Psychol. Sci.* **5** 457–469.

WALSH, C. G., RIBEIRO, J. D. and FRANKLIN, J. C. (2018). Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *J. Child Psychol. Psychiatry* **59** 1261–1270.

WEI, P., LU, Z. and SONG, J. (2015). Variable importance analysis: A comprehensive review. *Reliab. Eng. Syst. Saf.* **142** 399–432.

WILLIAMSON, B. (2023). lvimp: Perform Inference on Summaries of Longitudinal Algorithm-Agnostic Variable Importance R package version 0.0.0.9000.

WILLIAMSON, B. and FENG, J. (2020). Efficient nonparametric statistical inference on population feature importance using Shapley values. In *Proceedings of the 37th International Conference on Machine Learning*. *Proceedings of Machine Learning Research* **119** 10282–10291.

WILLIAMSON, B. D., GILBERT, P. B., CARONE, M. and SIMON, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics* **77** 9–22. MR4229718 https://doi.org/10.1111/biom.13392

WILLIAMSON, B. D., GILBERT, P. B., SIMON, N. R. and CARONE, M. (2023). A general framework for inference on algorithm-agnostic variable importance. *J. Amer. Statist. Assoc.* **118** 1645–1658. MR4646595 https://doi.org/10.1080/01621459.2021.2003200

WILLIAMSON, B. D., MOODIE, E. E., SIMON, G. E., ROSSOM, R. C. and SHORTREED, S. M. (2026). Supplement to "Inference on summaries of a model-agnostic longitudinal variable importance trajectory with application to suicide prevention." https://doi.org/10.1214/26-AOAS2186SUPPA, https://doi.org/10.1214/26-AOAS2186SUPPB

WOLOCK, C. J., WILLIAMSON, B. D., SHORTREED, S. M., SIMON, G. E., COLEMAN, K. J. et al. (2024). Importance of variables from different time frames for predicting self-harm using health system data. medRxiv 2024-04.

WU, J., GALANTER, N., SHORTREED, S. M. and MOODIE, E. E. M. (2022). Ranking tailoring variables for constructing individualized treatment rules: An application to schizophrenia. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **71** 309–330. MR4396911 https://doi.org/10.1111/rssc.12533

ZHENG, L., WANG, O., HAO, S., YE, C., LIU, M., XIA, M., SABO, A. N., MARKOVIC, L., STEARNS, F. et al. (2020). Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. *Transl. Psychiatry* **10** 72.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1364–1387
https://doi.org/10.1214/26-AOAS2153

TRANSPARENT SEQUENTIAL LEARNING AND MONITORING OF SPATIOTEMPORAL DISEASE INCIDENCE RATES

BY YUHANG ZHOU[a] AND PEIHUA QIU[b]

Department of Biostatistics, University of Florida, [a]*zhouyuhang@ufl.edu,* [b]*pqiu@ufl.edu*

Our society is under constant threat from outbreaks of various infectious diseases, such as COVID-19, Zika, and others. The recent COVID-19 pandemic has claimed millions of lives and caused devastating disruption to our daily routines. Prompt detection of outbreaks and effective disease surveillance are critical yet challenging, due to the complex spatiotemporal dynamics of infectious disease spread. Existing analytical tools often rely on restrictive assumptions, such as data independence and specific parametric distributions, that are rarely valid in practical scenarios. Additionally, effective disease surveillance demands sequential decision-making since decisions should be made or updated whenever new data become available. But many existing methods were designed for retrospective data observed in a prespecified time interval and would not be effective for disease surveillance. To address these limitations, we develop a cumulative sum (CUSUM) control chart for sequentially monitoring the evolution of spatiotemporal disease incidence rates. The new chart is constructed under the sequential learning framework that continuously incorporates new data in updating the estimated baseline model. Unlike traditional methods, the new method can capture complex data structure including spatiotemporal data variation and correlation. Numerical studies indicate that it achieves faster and more reliable detection of disease outbreaks compared to some traditional methods.

REFERENCES

APLEY, D. W. and TSUNG, F. (2002). The autoregressive T^2 chart for monitoring univariate autocorrelated processes. *J. Qual. Technol.* **34** 80–96.

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

BEST, N., RICHARDSON, S. and THOMSON, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Stat. Methods Med. Res.* **14** 35–59. MR2135924 https://doi.org/10.1191/0962280205sm388oa

BODNAR, R., BODNAR, T. and SCHMID, W. (2023). Sequential monitoring of high-dimensional time series. *Scand. J. Stat.* **50** 962–992. MR4630611 https://doi.org/10.1111/sjos.12607

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

CHEN, N., ZI, X. and ZOU, C. (2016). A distribution-free multivariate control chart. *Technometrics* **58** 448–459. MR3556613 https://doi.org/10.1080/00401706.2015.1049750

DIGGLE, P., ROWLINGSON, B. and SU, T. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics* **16** 423–434. MR2147534 https://doi.org/10.1002/env.712

EPSNEČNIKOV, V. A. (1969). Nonparametric estimation of a multidimensional probability density. *Theory Probab. Appl.* **14** 153–158. MR0250422

GEE, A. H., CHANG, J., GHOSH, J. and PAYDARFAR, D. (2018). Bayesian online changepoint detection of physiological transitions. In *Proceedings of the* 40*th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (*EMBC*) 45–48, Honolulu, HI, USA.

HANIGAN, I., HALL, G. and DEAR, K. B. (2006). A comparison of methods for calculating population exposure estimates of daily weather for health research. *Int. J. Health Geogr.* **5** 38.

HAWKINS, D. M. (1987). Self-starting cusums for location and scale. *Statistician* **36** 299–315.

HAWKINS, D. M., QIU, P. and KANG, C. W. (2003). The changepoint model for statistical process control. *J. Qual. Technol.* **35** 355–366.

Key words and phrases. Control charts, correlation, disease incidence rates, process monitoring, sequential learning, spatiotemporal data.

HEISTERKAMP, S. H., DEKKERS, A. L. M. and HEIJNE, J. C. M. (2006). Automated detection of infectious disease outbreaks: Hierarchical time series models. *Stat. Med.* **25** 4179–4196. MR2307584 https://doi.org/10.1002/sim.2674

JACQUEZ, G. M. (1996). A k nearest neighbor test for spacetime interaction. *Stat. Med.* **15** 1935–1949.

KLEINMAN, K., LAZARUS, R. and PLATT, R. (2004). A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *Amer. J. Epidemiol.* **159** 217–224.

KNOX, E. and BARTLETT, M. (1964). The detection of space-time interactions. *J. R. Stat. Soc., Ser. C* **13** 25–30.

KULLDORFF, M. (1997). A spatial scan statistic. *Comm. Statist. Theory Methods* **26** 1481–1496. MR1456844 https://doi.org/10.1080/03610929708831995

KULLDORFF, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *J. Roy. Statist. Soc. Ser. A* **164** 61–72. MR1819022 https://doi.org/10.1111/1467-985X.00186

LAWSON, A. B., BIGGERI, A. B., BOEHNING, D., LESAFFRE, E., VIEL, J.-F., CLARK, A., SCHLATTMANN, P. and DIVINO, F. (2000). Disease mapping models: An empirical evaluation. *Stat. Med.* **19** 2217–2241.

LAWSON, A. B. and KLEINMAN, K. (2005). *Spatial and Syndromic Surveillance for Public Health*. Wiley, New York.

MANTEL, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27** 209–220.

MARSHALL, J. B., SPITZNER, D. J. and WOODALL, W. H. (2007). Use of the local Knox statistic for the prospective monitoring of disease occurrences in space and time. *Stat. Med.* **26** 1579–1593. MR2359160 https://doi.org/10.1002/sim.2603

MURRAY, J. and COHEN, A. L. (2017). Infectious disease surveillance. In *International Encyclopedia of Public Health* (S. R. Quah, ed.) 222–229.

PELAT, C., BOËLLE, P.-Y., COWLING, B. J., CARRAT, F., FLAHAULT, A. and ANSART, S. (2007). Online detection and quantification of epidemics. *BMC Med. Inform. Decis. Mak.* **7** article 29.

QIU, P. (2005). *Image Processing and Jump Regression Analysis*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2111430 https://doi.org/10.1002/0471733156

QIU, P. (2014). *Introduction to Statistical Process Control*. CRC Press/CRC, Boca Raton, FL.

QIU, P. (2024). *Statistical Methods for Dynamic Disease Screening and Spatio-Temporal Disease Surveillance*. CRC Press/CRC, Boca Raton, FL.

QIU, P., LI, W. and LI, J. (2020). A new process control chart for monitoring short-range serially correlated data. *Technometrics* **62** 71–83. MR4058600 https://doi.org/10.1080/00401706.2018.1562988

QIU, P. and XIE, X. (2022). Transparent sequential learning for statistical process control of serially correlated data. *Technometrics* **64** 487–501. MR4506615 https://doi.org/10.1080/00401706.2021.1929493

QIU, P. and YANG, K. (2023). Spatiotemporal process monitoring using exponentially weighted spatial LASSO. *J. Qual. Technol.* **55** 163–180.

REIS, B. Y. and MANDL, K. D. (2003). Time series modeling for syndromic surveillance. *BMC Med. Inform. Decis. Mak.* **3**. article 2.

RODRIGUEZ AVELLANEDA, F., MATEU, J. and MORAGA, P. (2024). Estimating velocities of infectious disease spread through spatiotemporal log-Gaussian Cox point processes. arXiv preprint. Available at arXiv:2409.05036. https://doi.org/10.48550/arXiv.2409.05036

SONESSON, C. and BOCK, D. (2003). A review and discussion of prospective statistical surveillance in public health. *J. Roy. Statist. Soc. Ser. A* **166** 5–21. MR1973851 https://doi.org/10.1111/1467-985X.00256

STROUP, D., WHARTON, M., KAFADAR, K. and DEAN, A. (1999). Evaluation of a method for detecting aberrations in public health surveillance data. *Amer. J. Epidemiol.* **137** 373–380.

TAKAHASHI, K., KULLDORFF, M., TANGO, T. and YIH, K. (2008). A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *Int. J. Health Geogr.* **7** article 14.

WANG, Y., CHEN, X. and XUE, F. (2024). A review of Bayesian spatiotemporal models in spatial epidemiology. *ISPRS Int. J. Geo-Inf.* **13** 97. https://doi.org/10.3390/ijgi13030097

WANG, Z., GOEDHART, R. and ZWETSLOOT, I. M. (2023). Monitoring high-dimensional heteroscedastic processes using rank-based EWMA methods. *Comput. Ind. Eng.* **184** 109544. https://doi.org/10.1016/j.cie.2023.109544

WEINSTOCK, M. A. (1981). A generalized scan statistic test for the detection of clusters. *Int. J. Epidemiol.* **10** 289–293.

XIE, X. and QIU, P. (2023). Control charts for dynamic process monitoring with an application to air pollution surveillance. *Ann. Appl. Stat.* **17** 47–66. MR4539021 https://doi.org/10.1214/22-aoas1615

YAN, H., PAYNABAR, K. and SHI, J. (2018). Real-time monitoring of high-dimensional functional data streams via spatio-temporal smooth sparse decomposition. *Technometrics* **60** 181–197. MR3804247 https://doi.org/10.1080/00401706.2017.1346522

YANG, K. and QIU, P. (2018). Spatiotemporal incidence rate data analysis by nonparametric regression. *Stat. Med.* **37** 2094–2107. MR3802936 https://doi.org/10.1002/sim.7622

YANG, K. and QIU, P. (2019). Nonparametric estimation of the spatio-temporal covariance structure. *Stat. Med.* **38** 4555–4565. MR4010035 https://doi.org/10.1002/sim.8315

YANG, K. and QIU, P. (2020). Online sequential monitoring of spatiotemporal disease incidence rates. *IISE Trans.* **52** 1218–1233.

YU, Y., MADRID PADILLA, O. H., WANG, D. and RINALDO, A. (2023). A note on online change point detection. *Sequential Anal.* **42** 438–471. MR4678160 https://doi.org/10.1080/07474946.2023.2276170

ZHANG, J., KANG, Y., YANG, Y. and QIU, P. (2015). Statistical monitoring of the hand, foot and mouth disease in China. *Biometrics* **71** 841–850. MR3402620 https://doi.org/10.1111/biom.12301

ZHAO, Y., ZENG, D., HERRING, A. H., ISING, A., WALLER, A., RICHARDSON, D. and KOSOROK, M. R. (2011). Detecting disease outbreaks using local spatiotemporal methods. *Biometrics* **67** 1508–1517. MR2872402 https://doi.org/10.1111/j.1541-0420.2011.01585.x

ZHOU, H. and LAWSON, A. B. (2008). EWMA smoothing and Bayesian spatial modeling for health surveillance. *Stat. Med.* **27** 5907–5928. MR2597751 https://doi.org/10.1002/sim.3409

ZHOU, Y. and QIU, P. (2026). Supplement to "Transparent Sequential Learning and Monitoring of Spatiotemporal Disease Incidence Rates." https://doi.org/10.1214/26-AOAS2153SUPPA, https://doi.org/10.1214/26-AOAS2153SUPPB

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1388–1408
https://doi.org/10.1214/26-AOAS2172

MODELING TEMPORAL DEPENDENCE IN A SEQUENCE OF SPATIAL RANDOM PARTITIONS DRIVEN BY SPANNING TREE: AN APPLICATION TO MOSQUITO-BORNE DISEASES

BY JESSICA PAVANI[1,a], ROSANGELA H. LOSCHI[2,b] AND FERNANDO A. QUINTANA[3,c]

[1]*Department of Mathematics and Statistics, University of Calgary,* [a]*jessica.pavani@ucalgary.ca*
[2]*Departamento de Estatística, Universidade Federal de Minas Gerais,* [b]*loschi@est.ufmg.br*
[3]*Departamento de Estadística, Pontificia Universidad Católica de Chile,* [c]*quintana@uc.cl*

Time-dependent regionalization, or spatially restricted grouping, is a significant area of research focused on understanding the evolution of spatial clusters over time. In this study we adopt a probabilistic approach to regionalization, conceptualizing it as a random partition of geographic space at each time point, with the sequence of spatial partitions exhibiting time dependency. This methodology facilitates inference regarding the temporal dynamics of clusters. We employ a product partition prior for the random partitions at each time point, introducing temporal correlation among partitions through the temporal structure associated with prior cohesions. To explore partition search space effectively and ensure spatially constrained clustering, we utilize random spanning trees. This research is motivated by a pertinent applied problem: the identification of spatial and temporal patterns associated with mosquito-borne diseases. Given the overdispersion inherent in this type of data, we propose a spatiotemporal Poisson mixture model in which both mean and dispersion parameters vary according to spatiotemporal covariates. We apply the proposed model to analyze weekly reported cases of dengue from 2018 to 2023 in the Southeast region of Brazil. Additionally, we assess modeling performance using simulated data. Results indicate that our model is competitive in analyzing the temporal evolution of spatial clustering.

REFERENCES

ASSUNÇÃO, R. M., NEVES, M. C., CÂMARA, G. and FREITAS, C. C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *Int. J. Geogr. Inf. Sci.* **20** 797–811. https://doi.org/10.1080/13658810600665111

BARRETO-SOUZA, W. and SIMAS, A. B. (2016). General mixed Poisson regression models with varying dispersion. *Stat. Comput.* **26** 1263–1280. MR3538636 https://doi.org/10.1007/s11222-015-9601-6

BARRY, D. and HARTIGAN, J. A. (1993). A Bayesian analysis for change point problems. *J. Amer. Statist. Assoc.* **88** 309–319. MR1212493

CARON, F., NEISWANGER, W., WOOD, F., DOUCET, A. and DAVY, M. (2017). Generalized Pólya urn for time-varying Pitman–Yor processes. *J. Mach. Learn. Res.* **18** 27. MR3634894

CREMASCHI, A., CADONNA, A., GUGLIELMI, A. and QUINTANA, F. A. (2023). A change-point random partition model for large spatio-temporal datasets. Available at arXiv:2312.12396.

CRISCUOLO, T. L., ASSUNÇÃO, R. M., LOSCHI, R. H., MEIRA, W. JR. and CRUZ-REYES, D. (2023). Handling categorical features with many levels using a product partition model. *Ann. Appl. Stat.* **17** 786–814. MR4539053 https://doi.org/10.1214/22-aoas1651

DAHL, D., JOHNSON, D. and MÜLLER, P. (2020). salso: Search algorithms and loss functions for Bayesian clustering R package version 0.2.5. Available at https://cran.r-project.org/web/packages/salso/index.html.

DE IORIO, M., FAVARO, S., GUGLIELMI, A. and YE, L. (2023). Bayesian nonparametric mixture modeling for temporal dynamics of gender stereotypes. *Ann. Appl. Stat.* **17** 2256–2278. MR4637666 https://doi.org/10.1214/22-aoas1717

DOMBOWSKY, A. and DUNSON, D. B. (2025). Product centred Dirichlet processes for Bayesian multiview clustering. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **87** 1331–1352. MR4986678 https://doi.org/10.1093/jrsssb/qkaf021

Key words and phrases. Bayesian spatiotemporal clustering, correlated partitions, dengue, overdispersion, product partition model.

DUAN, L. L. and DUNSON, D. B. (2023). Bayesian spanning tree: Estimating the backbone of the dependence graph. *J. Mach. Learn. Res.* **24** 397. MR4733580

FRANKLINOS, L. HV., JONES, K. E., REDDING, D. W. and ABUBAKAR, I. (2019). The effect of global change on mosquito-borne disease. *Lancet Infect. Dis.* **19**. e302–e312. https://doi.org/10.1016/s1473-3099(19)30161-6.

FRANZOLINI, B., DE IORIO, M. and ERIKSSON, J. (2026). Conditional partial exchangeability: A probabilistic framework for multi-view clustering. *J. Amer. Statist. Assoc.* **1**. https://doi.org/10.1080/01621459.2025.2609381

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

GIAMPINO, A., NIPOTI, B., VANNUCCI, M. and GUINDANI, M. (2025). Local level dynamic random partition models for changepoint detection. *Bayesian Anal.* **1** 1–30. https://doi.org/10.1214/25-ba1560

GUTIÉRREZ, L., MENA, R. H. and RUGGIERO, M. (2016). A time dependent Bayesian nonparametric model for air quality analysis. *Comput. Statist. Data Anal.* **95** 161–175. MR3425946 https://doi.org/10.1016/j.csda.2015.10.002

HARTIGAN, J. A. (1990). Partition models. *Comm. Statist. Theory Methods* **19** 2745–2756. MR1088047 https://doi.org/10.1080/03610929008830345

HEGARTY, A. and BARRY, D. (2008). Bayesian disease mapping using product partition models. *Stat. Med.* **27** 3868–3893. MR2526613 https://doi.org/10.1002/sim.3253

HILBE, J. M. (2014). *Modeling Count Data*. Cambridge Univ. Press, Cambridge, UK.

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218. https://doi.org/10.1007/BF01908075

JARA, A., NIETO-BARAJAS, L. E. and QUINTANA, F. (2013). A time series model for responses on the unit interval. *Bayesian Anal.* **8** 723–740. MR3102232 https://doi.org/10.1214/13-BA844

JO, S., LEE, J., MÜLLER, P., QUINTANA, F. A. and TRIPPA, L. (2017). Dependent species sampling models for spatial density estimation. *Bayesian Anal.* **12** 379–406. MR3620738 https://doi.org/10.1214/16-BA1006

JUNGNICKEL, D. (2013). *Graphs, Networks and Algorithms*, 4th ed. *Algorithms and Computation in Mathematics* **5**. Springer, Heidelberg. MR2986014 https://doi.org/10.1007/978-3-642-32278-5

LUO, Z. T., SANG, H. and MALLICK, B. (2021). A Bayesian contiguous partitioning method for learning clustered latent variables. *J. Mach. Learn. Res.* **22** 37. MR4253730

LUO, Z. T., SANG, H. and MALLICK, B. (2024). A nonstationary soft partitioned Gaussian process model via random spanning trees. *J. Amer. Statist. Assoc.* **119** 2105–2116. MR4797926 https://doi.org/10.1080/01621459.2023.2249642

MARINHO, R. A., BESERRA, E. B., BEZERRA-GUSMÃO, M. A., PORTO, V. D. S., OLINDA, R. A. and DOS SANTOS, C. A. C. (2016). Effects of temperature on the life cycle, expansion, and dispersion of Aedes aegypti (Diptera: Culicidae) in three cities in Paraiba, Brazil. *J. Vector Ecol.* **41** 1–10. https://doi.org/10.1111/jvec.12187

MORAGA, P. (2019). *Geospatial Health Data*: *Modeling and Visualization with R-INLA and Shiny*. Chapman & Hall/CRC Biostatistics Series, Boca Raton, US.

NAPIER, G., LEE, D., ROBERTSON, C. and LAWSON, A. (2019). A Bayesian space-time model for clustering areal units based on their disease trends. *Biostatistics* **20** 681–697. MR4019725 https://doi.org/10.1093/biostatistics/kxy024

PAGE, G. L. and QUINTANA, F. A. (2016). Spatial product partition models. *Bayesian Anal.* **11** 265–298. MR3465813 https://doi.org/10.1214/15-BA971

PAGE, G. L., QUINTANA, F. A. and DAHL, D. B. (2022). Dependent modeling of temporal sequences of random partitions. *J. Comput. Graph. Statist.* **31** 614–627. MR4425090 https://doi.org/10.1080/10618600.2021.1987255

PAVANI, J., LOSCHI, R. H and QUINTANA, F. A (2026). Supplement to "Modeling temporal dependence in a sequence of spatial random partitions driven by spanning tree: an application to mosquito-borne diseases." https://doi.org/10.1214/26-AOAS2172SUPP

PAVANI, J. and QUINTANA, F. A. (2025). A Bayesian multivariate model with temporal dependence on random partition of areal data for mosquito-borne diseases. *Stat. Med.* **44** e10325. MR4860478 https://doi.org/10.1002/sim.10325

PERRAKIS, K., KARLIS, D., COOLS, M. and JANSSENS, D. (2015). Bayesian inference for transportation origin-destination matrices: The Poisson-inverse Gaussian and other Poisson mixtures. *J. Roy. Statist. Soc., Ser. A* **178** 271–296. MR3291771 https://doi.org/10.1111/rssa.12057

QUINTANA, F. A., LOSCHI, R. H. and PAGE, G. L. (2018). Bayesian product partition models. In *Wiley StatsRef*: *Statistics Reference Online* 1–15. https://doi.org/10.1002/9781118445112.stat08123

SARAIVA, E. F., VIGAS, V. P., FLESCH, M. V., GANNON, M. and DE BRAGANÇA PEREIRA, C. A. (2022). Modeling overdispersed Dengue data via Poisson inverse Gaussian regression model: A case study in the city of Campo Grande, MS, Brazil. *Entropy* **24** 1256. https://doi.org/10.3390/e24091256

TAM, E., DUNSON, D. B. and DUAN, L. L. (2025). Exact sampling of spanning trees via fast-forwarded random walks. *Biometrika* **112** asaf031. MR4926282 https://doi.org/10.1093/biomet/asaf031

TEIXEIRA, L. V., ASSUNÇÃO, R. M. and LOSCHI, R. H. (2015). A generative spatial clustering model for random data through spanning trees. In *IEEE International Conference on Data Mining* 997–1002. https://doi.org/10.1109/ICDM.2015.106

TEIXEIRA, L. V., ASSUNÇÃO, R. M. and LOSCHI, R. H. (2019). Bayesian space-time partitioning by sampling and pruning spanning trees. *J. Mach. Learn. Res.* **20** 85. MR3960939

ZHONG, R., CHACÓN-MONTALVÁN, E. and MORAGA, P. (2024). Bayesian spatial functional data clustering: Applications in disease surveillance. Available at arXiv:2407.12633v1.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1409–1428
https://doi.org/10.1214/26-AOAS2195

NEURAL POSTERIOR ESTIMATION FOR STOCHASTIC EPIDEMIC MODELING

BY PRAYAG CHATHA[1,a], FAN BU[2,c], JEFFREY REGIER[1,b], EVAN SNITKIN[3,d] AND JON ZELNER[4,e]

[1]*Department of Statistics, University of Michigan,* [a]*pchatha@umich.edu,* [b]*regier@umich.edu*
[2]*Department of Biostatistics, University of Michigan,* [c]*fbu@umich.edu*
[3]*Department of Microbiology and Immunology, University of Michigan,* [d]*esnitkin@umich.edu*
[4]*Department of Epidemiology & Center for Social Epidemiology and Population Health, University of Michigan,* [e]*jzelner@umich.edu*

Stochastic infectious disease models capture uncertainty in public health outcomes and have become increasingly popular in epidemiological practice. However, it is hard to calibrate realistic stochastic models to data due to the challenges of likelihood-based inference of unknown parameters. Stochastic epidemic models are nonlinear dynamical systems that may feature massive latent state spaces, resulting in computationally intractable likelihood densities. We develop an approach to calibrating large-scale epidemiological models using Neural Posterior Estimation, an emergent deep learning technique for simulation-based inference. In NPE, a neural network trained on simulated data learns to "invert" a stochastic simulator, returning a parametric approximation to the posterior distribution. Motivated by the problem of understanding transmission of carbapenem-resistant Klebsiella pneumoniae (CRKP), a major healthcare-associated infection, we propose a stochastic, discrete-time susceptible infected model. Through a realistic simulation experiment, we show that NPE produces accurate posterior estimates of unknown infection rates at a computational discount compared to Approximate Bayesian Computation. In an empirical study of CRKP transmission in a Chicago-area hospital, we use NPE to analyze spatial heterogeneity in patient-to-patient transmission risk.

REFERENCES

AMBROGIONI, L., GÜÇLÜ, U., BEREZUTSKAYA, J., BORNE, E., GÜÇLÜTÜRK, Y., HINNE, M., MARIS, E. and GERVEN, M. (2019). Forward amortized inference for likelihood-free variational marginalization. In *The* 22*nd International Conference on Artificial Intelligence and Statistics* 777–786. PMLR.

BAYDIN, A. G., SHAO, L., BHIMJI, W., HEINRICH, L., NADERIPARIZI, S., MUNK, A., LIU, J., GRAM-HANSEN, B., LOUPPE, G. et al. (2019). Efficient probabilistic inference in the quest for physics beyond the standard model. *Adv. Neural Inf. Process. Syst.* **32**.

BEAUMONT, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.* **41** 379–406.

BRETÓ, C., HE, D., IONIDES, E. L. and KING, A. A. (2009). Time series analysis via mechanistic models. *Ann. Appl. Stat.* **3** 319–348. MR2668710 https://doi.org/10.1214/08-AOAS201

BRITTON, T. (2010). Stochastic epidemic models: A survey. *Math. Biosci.* **225** 24–35. MR2642269 https://doi.org/10.1016/j.mbs.2010.01.006

BU, F., AIELLO, A. E., XU, J. and VOLFOVSKY, A. (2022). Likelihood-based inference for partially observed epidemics on dynamic networks. *J. Amer. Statist. Assoc.* **117** 510–526. MR4399102 https://doi.org/10.1080/01621459.2020.1790376

CAUCHEMEZ, S. and FERGUSON, N. M. (2011). Methods to infer transmission risk factors in complex outbreak data. *J. R. Soc. Interface* **9** 456–469.

CHAN, J., PERRONE, V., SPENCE, J., JENKINS, P., MATHIESON, S. and SONG, Y. (2018). A likelihood-free inference framework for population genetic data using exchangeable neural networks. *Adv. Neural Inf. Process. Syst.* **31**.

Key words and phrases. Epidemiology, simulation-based inference, deep learning.

CHATHA, P., BU, F., REGIER, J., SNITKIN, E. and ZELNER, J. (2026). Supplement to "Neural posterior estimation for stochastic epidemic modeling." https://doi.org/10.1214/26-AOAS2195SUPPA, https://doi.org/10.1214/26-AOAS2195SUPPB, https://doi.org/10.1214/26-AOAS2195SUPPC https://doi.org/10.1214/26-AOAS2195SUPPD https://doi.org/10.1214/26-AOAS2195SUPPE

CHOWELL, G. (2017). Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. *Infect. Dis. Model.* **2** 379–398.

CRANMER, K., BREHMER, J. and LOUPPE, G. (2020). The frontier of simulation-based inference. *Proc. Natl. Acad. Sci. USA* **117** 30055–30062. MR4263287 https://doi.org/10.1073/pnas.1912789117

DALEY, D. J. and GANI, J. (2001). *Epidemic Modelling: An Introduction. Cambridge Studies in Mathematical Biology* **15**. Cambridge Univ. Press, Cambridge.

DAX, M., GREEN, S. R., GAIR, J., MACKE, J. H., BUONANNO, A. and SCHÖLKOPF, B. (2021). Real-time gravitational wave science with neural posterior estimation. *Phys. Rev. Lett.* **127** 241103.

DOAN, T. N., KONG, D. C. M., KIRKPATRICK, C. M. J. and MCBRYDE, E. S. (2014). Optimizing hospital infection control: The role of mathematical modeling. *Infect. Control Hosp. Epidemiol.* **35** 1521–1530.

DUNSON, D. B. (2001). Commentary: Practical advantages of Bayesian analysis of epidemiologic data. *Amer. J. Epidemiol.* **153** 1222–1226.

ENDO, A., VAN LEEUWEN, E. and BAGUELIN, M. (2019). Introduction to particle Markov-chain Monte Carlo for disease dynamics modellers. *Epidemics* **29** 100363.

FINTZI, J., WAKEFIELD, J. and MININ, V. N. (2022). A linear noise approximation for stochastic epidemic models fit to partially observed incidence counts. *Biometrics* **78** 1530–1541. MR4534376 https://doi.org/10.1111/biom.13538

FOREMAN-MACKEY, D., HOGG, D. W., LANG, D. and GOODMAN, J. (2013). emcee: The MCMC hammer. *Publ. Astron. Soc. Pac.* **125** 306.

FRAZIER, D. T. and DROVANDI, C. (2021). Robust approximate Bayesian inference with synthetic likelihood. *J. Comput. Graph. Statist.* **30** 958–976. MR4356598 https://doi.org/10.1080/10618600.2021.1875839

FRAZIER, D. T., ROBERT, C. P. and ROUSSEAU, J. (2020). Model misspecification in approximate Bayesian computation: Consequences and diagnostics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 421–444. MR4084170 https://doi.org/10.1111/rssb.12356

HAN, J. H., LAPP, Z., BUSHMAN, F., LAUTENBACH, E., GOLDSTEIN, E. J., MATTEI, L., HOFSTAEDTER, C. E., KIM, D., NACHAMKIN, I. et al. (2019). Whole-genome sequencing to identify drivers of carbapenem-resistant Klebsiella pneumoniae transmission within and between regional long-term acute-care hospitals. *Antimicrob. Agents Chemother.* **63** 10–1128.

HAWKEN, S. E., YELIN, R. D., LOLANS, K., PIRANI, A., WEINSTEIN, R. A., LIN, M. Y., HAYDEN, M. K. and SNITKIN, E. S. (2022). Threshold-free genomic cluster detection to track transmission pathways in health-care settings: A genomic epidemiology analysis. *Lancet Microbe* **3** e652–e662.

HAYDEN, M. K., LIN, M. Y., LOLANS, K., WEINER, S., BLOM, D., MOORE, N. M., FOGG, L., HENRY, D., LYLES, R. et al. (2015). Prevention of colonization and infection by Klebsiella pneumoniae carbapenemase–producing Enterobacteriaceae in long-term acute-care hospitals. *Clin. Infect. Dis.* **60** 1153–1161.

HE, D., IONIDES, E. L. and KING, A. A. (2010). Plug-and-play inference for disease dynamics: Measles in large and small populations as a case study. *J. R. Soc. Interface* **7** 271–283.

HILBORN, R. and MANGEL, M. (2013). *The Ecological Detective: Confronting Models with Data*. Princeton Univ. Press, Princeton.

IONIDES, E. L., BHADRA, A., ATCHADÉ, Y. and KING, A. (2011). Iterated filtering. *Ann. Statist.* **39** 1776–1802. MR2850220 https://doi.org/10.1214/11-AOS886

IONIDES, E. L., BRETÓ, C. and KING, A. A. (2006). Inference for nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* **103** 18438–18443.

KAO, R. R., HAYDON, D. T., LYCETT, S. J. and MURCIA, P. R. (2014). Supersize me: How whole-genome sequencing and big data are transforming epidemiology. *Trends Microbiol.* **22** 282–291.

KING, A. (2017). Simulation of stochastic dynamic models. Available at https://kingaa.github.io/clim-dis/. Accessed: 2024-05-12.

KLINKENBERG, D., BACKER, J. A., DIDELOT, X., COLIJN, C. and WALLINGA, J. (2017). Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput. Biol.* **13** e1005495.

LIU, R., MCAULIFFE, J. D., REGIER, J. and LSST DARK ENERGY SCIENCE COLLABORATION (2023). Variational inference for deblending crowded starfields. *J. Mach. Learn. Res.* **24** Paper No. [179]. MR4633568

LLEDO, W., HERNANDEZ, M., LOPEZ, E., MOLINARI, O., SOTO, R., HERNANDEZ, E., SANTIAGO, N., FLORES, M., VAZQUEZ, G. et al. (2009). Guidance for control of infections with carbapenem-resistant or carbapenemase-producing Enterobacteriaceae in acute care facilities. *Morb. Mort. Wkly. Rep.* **58**.

LOSHCHILOV, I., HUTTER, F. et al. (2017). Fixing weight decay regularization in ADAM. arXiv Preprint. Available at arXiv:1711.05101.

LUECKMANN, J.-M., BOELTS, J., GREENBERG, D., GONCALVES, P. and MACKE, J. (2021). Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics* 343–351. PMLR.

LUECKMANN, J.-M., GONCALVES, P. J., BASSETTO, G., ÖCAL, K., NONNENMACHER, M. and MACKE, J. H. (2017). Flexible statistical inference for mechanistic models of neural dynamics. *Adv. Neural Inf. Process. Syst.* **30**.

MACDONALD, M. G. (1957). *The Epidemiology and Control of Malaria*. Oxford Univ. Press, London.

MACKAY, D. J. (1992). Bayesian interpolation. *Neural Comput.* **4** 415–447.

MADDEN, W. G., JIN, W., LOPMAN, B., ZUFLE, A., DALZIEL, B., METCALF, C. J. E., GRENFELL, B. T. and LAU, M. S. (2024). Deep neural networks for endemic measles dynamics: Comparative analysis and integration with mechanistic models. *PLoS Comput. Biol.* **20** e1012616.

MCKINLEY, T. J., VERNON, I., ANDRIANAKIS, I., MCCREESH, N., OAKLEY, J. E., NSUBUGA, R. N., GOLDSTEIN, M. and WHITE, R. G. (2018). Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models. *Statist. Sci.* **33** 4–18. MR3757500 https://doi.org/10.1214/17-STS618

MCNAMARA, D., LOPER, J. and REGIER, J. (2024). Globally convergent variational inference. *Adv. Neural Inf. Process. Syst.* **37** 18557–18592.

MINTER, A. and RETKUTE, R. (2019). Approximate Bayesian computation for infectious disease modelling. *Epidemics* **29** 100368.

PAPAMAKARIOS, G. (2019). Neural density estimation and likelihood-free inference. arXiv Preprint. Available at arXiv:1910.13233.

PAPAMAKARIOS, G. and MURRAY, I. (2016). Fast ε-free inference of simulation models with Bayesian conditional density estimation. *Adv. Neural Inf. Process. Syst.* **29**.

PAPAMAKARIOS, G., NALISNICK, E., REZENDE, D. J., MOHAMED, S. and LAKSHMINARAYANAN, B. (2021). Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.* **22** Paper No. 57. MR4253750

RADEV, S. T., GRAW, F., CHEN, S., MUTTERS, N. T., EICHEL, V. M., BÄRNIGHAUSEN, T. and KÖTHE, U. (2021). OutbreakFlow: Model-based Bayesian inference of disease outbreak dynamics with invertible neural networks and its application to the COVID-19 pandemics in Germany. *PLoS Comput. Biol.* **17** e1009472.

ROSS, R. (1911). *The Prevention of Malaria*. Murray, London.

SISSON, S. A., FAN, Y. and BEAUMONT, M. (2018). *Handbook of Approximate Bayesian Computation*. CRC Press.

SPURIO MANCINI, A., DOCHERTY, M., PRICE, M. and MCEWEN, J. (2023). Bayesian model comparison for simulation-based inference. *RAS Tech. Instrum.* **2** 710–722.

TEJERO-CANTERO, A., BOELTS, J., DEISTLER, M., LUECKMANN, J.-M., DURKAN, C., GONÇALVES, P. J., GREENBERG, D. S. and MACKE, J. H. (2020). sbi: A toolkit for simulation-based inference. *J. Open Sour. Softw.* **5** 2505.

TONI, T. and STUMPF, M. P. (2010). Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* **26** 104–110.

TONI, T., WELCH, D., STRELKOWA, N., IPSEN, A. and STUMPF, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6** 187–202.

VASIST, M., ROZET, F., ABSIL, O., MOLLIÈRE, P., NASEDKIN, E. and LOUPPE, G. (2023). Neural posterior estimation for exoplanetary atmospheric retrieval. *Astron. Astrophys.* **672** A147.

WARD, D., CANNON, P., BEAUMONT, M., FASIOLO, M. and SCHMON, S. (2022). Robust neural posterior estimation and statistical model criticism. *Adv. Neural Inf. Process. Syst.* **35** 33845–33859.

WOOD, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466** 1102–1104.

ZELNER, J. and EISENBERG, M. (2022). Rapid response modeling of SARS-CoV-2 transmission. *Science* **376** 579–580.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1429–1451
https://doi.org/10.1214/26-AOAS2154

A NOVEL BAYESIAN FRAMEWORK UNCOVERING BRAIN CONNECTIVITY-TO-SHAPE RELATIONSHIP IN PRECLINICAL ALZHEIMER'S DISEASE

BY SHENGXIAN DING[1,a], EMILY JOHNS[1,b], ANTON ORLICHENKO[1,c], CAROLYN FREDERICKS[2,e] AND YIZE ZHAO[1,d]

[1]*Department of Biostatistics, Yale University,* [a]*naomi.ding@yale.edu,* [b]*emily.johns@yale.edu,* [c]*anton.orlichenko@yale.edu,* [d]*yize.zhao@yale.edu*

[2]*Department of Neurology, Yale University,* [e]*carolyn.fredericks@yale.edu*

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by amyloid-beta plaques and tau tangles, with significant pathological changes occurring in subcortical brain regions. While previous research has focused primarily on volumetric reductions in areas, such as the hippocampus, thalamus, and caudate, emerging evidence suggests that their fine-grained shape deformations may offer greater sensitivity to early disease pathology. Moreover, understanding how these shape alterations influence brain functional connectivity (FC) networks could provide critical insights into the neurobiological mechanisms underlying the progression of AD. In this context we propose a novel statistical approach, the Connectivity-on-Shape Regression (COSR) model, designed to investigate the spatially varying impact of brain subcortical shape on FC, accounting for the intrinsic modularity of functional networks. Under a Bayesian framework, COSR employs a relaxed-thresholded Gaussian process prior model to promote feature selection and integrates a stochastic block model to capture the unknown modular organization of FC. To facilitate the practical application of COSR with vertex-level shape measurements, we develop a computationally efficient variational inference approach to achieve posterior inference. Extensive simulations demonstrate the superiority of COSR over existing alternatives in accurately uncovering connectivity-to-shape associations and identifying neurobiological signals. Applying COSR to data from the Anti-Amyloid Treatment in Asymptomatic Alzheimer's study, we discover meaningful neural structural-functional relationships in amyloid-positive individuals, highlighting the potentially complex interplay between structural and functional brain alterations during this crucial preclinical stage of AD.

REFERENCES

AGGLETON, J. P., PRALUS, A., NELSON, A. J. and HORNBERGER, M. (2016). Thalamic pathology and memory loss in early Alzheimer's disease: Moving the focus from the medial temporal lobe to Papez circuit. *Brain* **139** 1877–1890.

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

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. *Information Science and Statistics*. Springer, New York. MR2247587 https://doi.org/10.1007/978-0-387-45528-0

BLANKEN, A. E., HURTZ, S., ZAROW, C., BIADO, K., HONARPISHEH, H., SOMME, J., BROOK, J., TUNG, S., KRAFT, E. et al. (2017). Associations between hippocampal morphometry and neuropathologic markers of Alzheimer's disease using 7 T MRI. *NeuroImage Clin.* **15** 56–61.

BROVELLI, A., NAZARIAN, B., MEUNIER, M. and BOUSSAOUD, D. (2011). Differential roles of caudate nucleus and putamen during instrumental learning. *NeuroImage* **57** 1580–1590.

BRYCE, N. V., FLOURNOY, J. C., MOREIRA, J. F. G., ROSEN, M. L., SAMBOOK, K. A., MAIR, P. and MCLAUGHLIN, K. A. (2021). Brain parcellation selection: An overlooked decision point with meaningful effects on individual differences in resting-state functional connectivity. *NeuroImage* **243** 118487.

Key words and phrases. Alzheimer's disease, functional connectivity, Gaussian process, shape analysis, variational inference.

BZDOK, D., LAIRD, A. R., ZILLES, K., FOX, P. T. and EICKHOFF, S. B. (2013). An investigation of the structural, connectional, and functional subspecialization in the human amygdala. *Hum. Brain Mapp.* **34** 3247–3266.

CHAPLEAU, M., BEDETTI, C., DEVENYI, G. A., SHELDON, S., ROSEN, H. J., MILLER, B. L., GORNO-TEMPINI, M. L., CHAKRAVARTY, M. M. and BRAMBATI, S. M. (2020). Deformation-based shape analysis of the hippocampus in the semantic variant of primary progressive aphasia and Alzheimer's disease. *NeuroImage Clin.* **27** 102305.

DE VOOGD, L. D., KLUMPERS, F., FERNÁNDEZ, G. and HERMANS, E. J. (2017). Intrinsic functional connectivity between amygdala and hippocampus during rest predicts enhanced memory under stress. *Psychoneuroendocrinology* **75** 192–202.

DELGORIO, P. L., HISCOX, L. V., DAUGHERTY, A. M., SANJANA, F., MCILVAIN, G., POHLIG, R. T., MCGARRY, M. D., MARTENS, C. R., SCHWARB, H. et al. (2022). Structure–function dissociations of human hippocampal subfield stiffness and memory performance. *J. Neurosci.* **42** 7957–7968.

DING, S., JOHNS, E., ORLICHENKO, A., FREDERICKS, C. and ZHAO, Y. (2026). Supplement to "A novel Bayesian framework uncovering brain connectivity-to-shape relationship in preclinical Alzheimer's disease." https://doi.org/10.1214/26-AOAS2154SUPPA, https://doi.org/10.1214/26-AOAS2154SUPPB

FILIPPI, M. and AGOSTA, F. (2011). Structural and functional network connectivity breakdown in Alzheimer's disease studied with magnetic resonance imaging techniques. *J. Alzheimer's Dis.* **24** 455–474.

GREENE, A. S., GAO, S., SCHEINOST, D. and CONSTABLE, R. T. (2018). Task-induced brain state manipulation improves prediction of individual traits. *Nat. Commun.* **9** 2807.

GREICIUS, M. D., SRIVASTAVA, G., REISS, A. L. and MENON, V. (2004). Default-mode network activity distinguishes Alzheimer's disease from healthy aging: Evidence from functional MRI. *Proc. Natl. Acad. Sci. USA* **101** 4637–4642.

HAMPTON, O. L., BUCKLEY, R. F., MANNING, L. K., SCOTT, M. R., PROPERZI, M. J., PEÑA-GÓMEZ, C., JACOBS, H. I. L., CHHATWAL, J. P., JOHNSON, K. A. et al. (2020). Resting-state functional connectivity and amyloid burden influence longitudinal cortical thinning in the default mode network in preclinical Alzheimer's disease. *NeuroImage Clin.* **28** 102407.

HARI, E., ULASOGLU-YILDIZ, C., KURT, E., BAYRAM, A., GURVIT, H. and DEMIRALP, T. (2024). Volumetric and functional connectivity changes of the thalamic nuclei in different stages of Alzheimer's disease. *Clin. Neurophysiol.* **165** 127–137.

HE, Y., WANG, J., WANG, L., CHEN, Z. J., YAN, C., YANG, H., TANG, H., ZHU, C., GONG, Q. et al. (2009). Uncovering intrinsic modular organization of spontaneous brain activity in humans. *PLoS ONE* **4** e5226.

HETT, K., TA, V.-T., CATHELINE, G., TOURDIAS, T., MANJÓN, J. V. and COUPÉ, P. (2019). Multimodal hippocampal subfield grading for Alzheimer's disease classification. *Sci. Rep.* **9** 13845.

HO, T. C., GUTMAN, B., POZZI, E., GRABE, H. J., HOSTEN, N., WITTFELD, K., VÖLZKE, H., BAUNE, B., DANNLOWSKI, U. et al. (2022). Subcortical shape alterations in major depressive disorder: Findings from the ENIGMA major depressive disorder working group. *Hum. Brain Mapp.* **43** 341–351.

HORIEN, C., SHEN, X., SCHEINOST, D. and CONSTABLE, R. T. (2019). The individual functional connectome is unique and stable over months to years. *NeuroImage* **189** 676–687.

HWANG, K., BERTOLERO, M. A., LIU, W. B. and D'ESPOSITO, M. (2017). The human thalamus is an integrative hub for functional brain networks. *J. Neurosci.* **37** 5594–5607.

JI, J. L., SPRONK, M., KULKARNI, K., REPOVŠ, G., ANTICEVIC, A. and COLE, M. W. (2019). Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage* **185** 35–57.

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

KERNBACH, J. M., YEO, B. T., SMALLWOOD, J., MARGULIES, D. S., THIEBAUT DE SCHOTTEN, M., WALTER, H., SABUNCU, M. R., HOLMES, A. J., GRAMFORT, A. et al. (2018). Subspecialization within default mode nodes characterized in 10,000 UK biobank participants. *Proc. Natl. Acad. Sci. USA* **115** 12295–12300.

LAANSMA, M. A., ZHAO, Y., VAN HEESE, E. M., BRIGHT, J. K., OWENS-WALTON, C., AL-BACHARI, S., ANDERSON, T. J., ASSOGNA, F., VAN BALKOM, T. D. et al. (2024). A worldwide study of subcortical shape as a marker for clinical staging in Parkinson's disease. *npj Parkinson's Dis.* **10** 223.

LEE, E.-C., KANG, J. M., SEO, S., SEO, H.-E., LEE, S.-Y., PARK, K. H., NA, D. L., NOH, Y. and SEONG, J.-K. (2020). Association of subcortical structural shapes with tau, amyloid, and cortical atrophy in early-onset and late-onset Alzheimer's disease. *Front. Aging Neurosci.* **12** 563559.

LI, J., CAO, D., YU, S., XIAO, X., IMBACH, L., STIEGLITZ, L., SARNTHEIN, J. and JIANG, T. (2023). Functional specialization and interaction in the amygdala-hippocampus circuit during working memory processing. *Nat. Commun.* **14** 2921.

LI, J., GONG, Y. and TANG, X. (2017). Hierarchical subcortical sub-regional shape network analysis in Alzheimer's disease. In *Neuroscience* **366** 70–83. Elsevier, Amsterdam.

MAKALIC, E. and SCHMIDT, D. F. (2016). High-dimensional Bayesian regularised regression with the BayesReg package. Preprint. Available at arXiv:1611.06649.

MENACHER, A., NICHOLS, T. E., JOHNSON, T. D. and KANG, J. (2024). Scalable scalar-on-image cortical surface regression with a relaxed-thresholded Gaussian process prior. Preprint. Available at arXiv:2403.13628.

MUELLER, S. G., CHAO, L., BERMAN, B. and WEINER, M. W. (2011). Evidence for functional specialization of hippocampal subfields detected by MR subfield volumetry on high resolution images at 4 T. *NeuroImage* **56** 851–857.

MÜLLER, M. J., GREVERUS, D., DELLANI, P. R., WEIBRICH, C., WILLE, P. R., SCHEURICH, A., STOETER, P. and FELLGIEBEL, A. (2005). Functional implications of hippocampal volume and diffusivity in mild cognitive impairment. *NeuroImage* **28** 1033–1042.

NAJAFI, M., MCMENAMIN, B. W., SIMON, J. Z. and PESSOA, L. (2016). Overlapping communities reveal rich structure in large-scale brain networks during rest and task conditions. *NeuroImage* **135** 92–106.

QIU, A., BROWN, T., FISCHL, B., MA, J. and MILLER, M. I. (2010). Atlas generation for subcortical and ventricular structures with its applications in shape analysis. *IEEE Trans. Image Process.* **19** 1539–1547. MR2814625 https://doi.org/10.1109/TIP.2010.2042099

RAHAYEL, S., BOCTI, C., SÉVIGNY DUPONT, P., JOANNETTE, M., LAVALLÉE, M. M., NIKELSKI, J., CHERTKOW, H. and JOUBERT, S. (2019). Subcortical amyloid load is associated with shape and volume in cognitively normal individuals. *Hum. Brain Mapp.* **40** 3951–3965.

RANASINGHE, P. and MAPA, M. S. (2024). Functional connectivity and cognitive decline: A review of rs-fMRI, EEG, MEG, and graph theory approaches in aging and dementia. *Explor. Med.* **5** 797–821.

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

SCHAEFER, A., KONG, R., GORDON, E. M., LAUMANN, T. O., ZUO, X.-N., HOLMES, A. J., EICKHOFF, S. B. and YEO, B. T. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* **28** 3095–3114.

SEELEY, W. W., CRAWFORD, R. K., ZHOU, J., MILLER, B. L. and GREICIUS, M. D. (2009). Neurodegenerative diseases target large-scale human brain networks. *Neuron* **62** 42–52.

SEOANE, S., VAN DEN HEUVEL, M., ACEBES, Á. and JANSSEN, N. (2024). The subcortical default mode network and Alzheimer's disease: A systematic review and meta-analysis. *Brain Commun.* **6** fcae128.

SHAPIRA, L., SHAMIR, A. and COHEN-OR, D. (2008). Consistent mesh partitioning and skeletonisation using the shape diameter function. *Vis. Comput.* **24** 249–259.

SHI, J., THOMPSON, P. M., GUTMAN, B. and WANG, Y. (2013). Surface fluid registration of conformal representation: Application to detect disease burden and genetic influence on hippocampus. *NeuroImage* **78** 111–134.

SPERLING, R. A., DONOHUE, M. C., RAMAN, R., SUN, C.-K., YAARI, R., HOLDRIDGE, K., SIEMERS, E., JOHNSON, K. A., AISEN, P. S. et al. (2020). Association of factors with elevated amyloid burden in clinically normal older individuals. *JAMA Neurol.* **77** 735–745.

STYNER, M., OGUZ, I., XU, S., BRECHBÜHLER, C., PANTAZIS, D., LEVITT, J. J., SHENTON, M. E. and GERIG, G. (2006). Framework for the statistical shape analysis of brain structures using SPHARM-PDM. *Insight J.* **1071** 242.

SUN, Z., XU, W., LI, T., KANG, J., ALANIS-LOBATO, G. and ZHAO, Y. (2025). Bayesian thresholded modeling for integrating brain node and network predictors. *Biostatistics* **26** Paper No. kxae048, 16. MR4848685 https://doi.org/10.1093/biostatistics/kxae048

THOMPSON, P. M., HAYASHI, K. M., DE ZUBICARAY, G. I., JANKE, A. L., ROSE, S. E., SEMPLE, J., HONG, M. S., HERMAN, D. H., GRAVANO, D. et al. (2004). Mapping hippocampal and ventricular change in Alzheimer disease. *NeuroImage* **22** 1754–1766.

THOMPSON, P. M., STEIN, J. L., MEDLAND, S. E., HIBAR, D. P., VASQUEZ, A. A., RENTERIA, M. E., TORO, R., JAHANSHAD, N., SCHUMANN, G. et al. (2014). The ENIGMA consortium: Large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav.* **8** 153–182.

TIAN, Y., MARGULIES, D. S., BREAKSPEAR, M. and ZALESKY, A. (2020). Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nat. Neurosci.* **23** 1421–1432.

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

WANG, Y., SONG, Y., RAJAGOPALAN, P., AN, T., LIU, K., CHOU, Y.-Y., GUTMAN, B., TOGA, A. W. and THOMPSON, P. M. (2011). Surface-based TBM boosts power to detect disease effects on the brain: An N = 804 ADNI study. *NeuroImage* **56** 1993–2010.

WHITEMAN, A. S., JOHNSON, T. D. and KANG, J. (2024). Bayesian inference for group-level cortical surface image-on-scalar regression with Gaussian process priors. *Biometrics* **80** Paper No. ujae116, 11. MR4896896 https://doi.org/10.1093/biomtc/ujae116

WIG, G. S. (2017). Segregated systems of human brain networks. *Trends Cogn. Sci.* **21** 981–996.

YEO, B. T., KRIENEN, F. M., SEPULCRE, J., SABUNCU, M. R., LASHKARI, D., HOLLINSHEAD, M., ROFFMAN, J. L., SMOLLER, J. W., ZÖLLEI, L. et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.*

YU, M., LINN, K. A., COOK, P. A., PHILLIPS, M. L., MCINNIS, M., FAVA, M., TRIVEDI, M. H., WEISSMAN, M. M., SHINOHARA, R. T. et al. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Hum. Brain Mapp.* **39** 4213–4227.

ZHANG, A. Y. and ZHOU, H. H. (2020). Theoretical and computational guarantees of mean field variational inference for community detection. *Ann. Statist.* **48** 2575–2598. MR4152113 https://doi.org/10.1214/19-AOS1898

ZHANG, Z., WU, Y., XIONG, D., IBRAHIM, J. G., SRIVASTAVA, A. and ZHU, H. (2022). LESA: Longitudinal Elastic Shape Analysis of Brain Subcortical Structures. J. Amer. Statist. Assoc. 1–27.

ZHU, H., FAN, J. and KONG, L. (2014). Spatially varying coefficient model for neuroimaging data with jump discontinuities. *J. Amer. Statist. Assoc.* **109** 1084–1098. MR3265682 https://doi.org/10.1080/01621459.2014.881742

ZHU, H., KHONDKER, Z., LU, Z. and IBRAHIM, J. G. (2014). Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *J. Amer. Statist. Assoc.* **109** 977–990. MR3265670 https://doi.org/10.1080/01621459.2014.923775

SAMTANI, M. N., FARNUM, M., LOBANOV, V., YANG, E., RAGHAVAN, N., DIBERNARDO, A., NARAYAN, V. and INITIATIVE, A. T. A. D. N. (2012). An improved model for disease progression in patients from the Alzheimer's Disease Neuroimaging Initiative. *J. Clin. Pharmacol.* **52** 629–644. https://onlinelibrary.wiley.com/doi/pdf/10.1177/0091270011405497.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1452–1476
https://doi.org/10.1214/26-AOAS2143

HIGH-DIMENSIONAL LOCALLY STATIONARY MODELS FOR IDENTIFYING NEUROIMAGING BIOMARKERS IN AUTISM SPECTRUM DISORDER

BY ZICHANG XIANG[1,a], RAANJU SUNDARARAJAN[1,b] AND HERNANDO OMBAO[2,c]

[1]*Department of Statistics and Data Science, Southern Methodist University,* [a]*zichangx@smu.edu,* [b]*rsundararajan@smu.edu*
[2]*Statistics Program, King Abdullah University of Science and Technology,* [c]*hernando.ombao@kaust.edu.sa*

Understanding functional connectivity patterns in autism spectrum disorder (ASD) remains an unsettled scientific problem with existing literature pointing to evidence of both over- and underconnectivity. Resting-state fMRI data is a popular modality with rich spatiotemporal information but poses significant modeling challenges. This work presents a new method for identifying neuroimaging biomarkers in ASD using high-dimensional resting-state fMRI data. The proposed model is flexible in that it can accommodate multiple features in fMRI time series data such as stationary, nonstationary, high-dimensional, Gaussian and non-Gaussian. Statistical inference is carried out by constructing an approximate Whittle likelihood that stems from a frequency domain factor model. The modeling approach also enables estimation of frequency-specific functional connectivity matrices. Two group comparisons (ASD vs. healthy control) are carried out by finding differences in the distributions and means of the number of edges in the functional connectivity matrices. These two group comparisons via testing are achieved using multiple candidate discrete and zero-inflated discrete distributions. Using our proposed approach led to interesting results that indicate altered functional connectivity in ASD with varying patterns witnessed across different age groups, resting-state networks and frequencies.

REFERENCES

ACHARD, S., SALVADOR, R., WHITCHER, B., SUCKLING, J. and BULLMORE, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *J. Neurosci*. **26** 63–72. https://doi.org/10.1523/JNEUROSCI.3874-05.2006

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

BERNARD, F., LEMEE, J.-M., MAZERAND, E., LEIBER, L.-M., MENEI, P. and TER MINASSIAN, A. (2020). The ventral attention network: The mirror of the language network in the right brain hemisphere. *J. Anat*. **237** 632–642. https://doi.org/10.1111/joa.13223

BI, X., ZHAO, Z., CHEN, Z., CHEN, N. and LI, S. (2018). A study on resting-state functional connectivity in the dorsal attention network of Autism Spectrum Disorder. *Front. Physiol*. **9** 475. https://doi.org/10.3389/fphys.2018.00475

BRILLINGER, D. (2001). *Time Series*. *Society for Industrial and Applied Mathematics*. https://doi.org/10.1137/1.9780898719246

CHENJI, S., JHA, S., LEE, D., BROWN, M., SERES, P., MAH, D. and KALRA, S. (2016). Investigating default mode and sensorimotor network connectivity in amyotrophic lateral sclerosis. *PLoS ONE* **11** 1–14. https://doi.org/10.1371/journal.pone.0157443

CHIEN, H.-Y., LIN, H.-Y., LAI, M.-C., GAU, S. S.-F. and TSENG, W.-Y. I. (2015). Hyperconnectivity of the right posterior temporo-parietal junction predicts social difficulties in boys with Autism Spectrum Disorder. *Autism Res*. **8** 427–441. https://doi.org/10.1002/aur.1457

CONG, J., ZHUANG, W., LIU, Y., YIN, S., JIA, H., YI, C., CHEN, K., XUE, K., LI, F. et al. (2023). Altered default mode network causal connectivity patterns in autism spectrum disorder revealed by Liang information flow analysis. *Hum. Brain Mapp*. **44** 2279–2293. https://doi.org/10.1002/hbm.26209

Key words and phrases. fMRI time series, frequency domain, factor model, resting-state network, Autism Spectrum Disorder.

CONTRERAS-CRISTÁN, A., GUTIÉRREZ-PEÑA, E. and WALKER, S. G. (2006). A note on Whittle's likelihood. *Comm. Statist. Simulation Comput.* **35** 857–875. https://doi.org/10.1080/03610910600880203

CRADDOCK, C., BENHAJALI, Y., CHU, C., CHOUINARD, F., EVANS, A., JAKAB, A., KHUNDRAKPAM, B. S., LEWIS, J. D., LI, Q. et al. (2013). The Neuro Bureau preprocessing initiative: Open sharing of preprocessed neuroimaging data and derivatives. *Front. Neuroinform.* **7**.

CRIBBEN, I., HARALDSDOTTIR, R., ATLAS, L. Y., WAGER, T. D. and LINDQUIST, M. A. (2012). Dynamic connectivity regression: Determining state-related changes in brain connectivity. *NeuroImage* **61** 907–920. https://doi.org/10.1016/j.neuroimage.2012.03.070

CRIBBEN, I., WAGER, T. and LINDQUIST, M. (2013). Detecting functional connectivity change points for single-subject fMRI data. *Front. Comput. Neurosci.* **7** 143. https://doi.org/10.3389/fncom.2013.00143

DAHLHAUS, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25** 1–37.

DAHLHAUS, R. (2012). 13—locally stationary processes. In *Time Series Analysis*: *Methods and Applications* (S. S. R. Tata Subba Rao and C. R. Rao, eds.). *Handbook of Statistics* **30** 351–413. Elsevier, Amsterdam. https://doi.org/10.1016/B978-0-444-53858-1.00013-2

DEVITT, N. M., GALLAGHER, L. and REILLY, R. B. (2015). Autism Spectrum Disorder (ASD) and Fragile X Syndrome (FXS): Two overlapping disorders reviewed through electroencephalography—what can be interpreted from the available information? *Brain Sci.* **5** 92–117.

EMBLETON, J., KNIGHT, M. I. and OMBAO, H. (2022). Multiscale spectral modelling for nonstationary time series within an ordered multiple-trial experiment. *Ann. Appl. Stat.* **16** 2774–2803. https://doi.org/10.1214/22-AOAS1614

FIECAS, M. and OMBAO, H. (2016). Modeling the evolution of dynamic brain processes during an associative learning experiment. *J. Amer. Statist. Assoc.* **111** 1440–1453. https://doi.org/10.1080/01621459.2016.1165683

GREICIUS, M. D. and MENON, V. (2004). Default-mode activity during a passive sensory task: Uncoupled from deactivation but impacting activation. *J. Cogn. Neurosci.* **16** 1484–1492. https://doi.org/10.1162/0898929042568532

HU, Z. and PRADO, R. (2023). Fast Bayesian inference on spectral analysis of multivariate stationary time series. *Comput. Statist. Data Anal.* **178** 107596. https://doi.org/10.1016/j.csda.2022.107596

HULL, J. V., DOKOVNA, L. B., JACOKES, Z. J., TORGERSON, C. M., IRIMIA, A. and VAN HORN, J. D. (2017). Resting-state functional connectivity in Autism Spectrum Disorders: A review. *Front. Psychiatry* **7** 205. https://doi.org/10.3389/fpsyt.2016.00205

KENNEDY, D. P. and COURCHESNE, E. (2008a). The intrinsic functional organization of the brain is altered in autism. *NeuroImage* **39** 1877–1885. https://doi.org/10.1016/j.neuroimage.2007.10.052

KENNEDY, D. P. and COURCHESNE, E. (2008b). The intrinsic functional organization of the brain is altered in autism. *NeuroImage* **39** 1877–1885. https://doi.org/10.1016/j.neuroimage.2007.10.052

KEOWN, C. L., SHIH, P., NAIR, A., PETERSON, N., MULVEY, M. E. and MÜLLER, R.-A. (2013). Local functional overconnectivity in posterior brain regions is associated with symptom severity in Autism Spectrum Disorders. *Cell Rep.* **5** 567–572. https://doi.org/10.1016/j.celrep.2013.10.003

KIRCH, C., EDWARDS, M. C., MEIER, A. and MEYER, R. (2019). Beyond Whittle: Nonparametric correction of a parametric likelihood with a focus on Bayesian time series analysis. *Bayesian Anal.* **14** 1037–1073. https://doi.org/10.1214/18-BA1126

LA, C., YOUNG, B. M., GARCIA-RAMOS, C., NAIR, V. A. and PRABHAKARAN, V. (2014). Chapter twenty – characterizing recovery of the human brain following stroke: Evidence from fMRI studies. In *Imaging of the Human Brain in Health and Disease* (P. Seeman and B. Madras, eds.) 485–506. Academic Press, Boston, MA. https://doi.org/10.1016/B978-0-12-418677-4.00020-8

LI, R., CHEN, K., FLEISHER, A. S., REIMAN, E. M., YAO, L. and WU, X. (2011). Large-scale directional connections among multi resting-state neural networks in human brain: A functional MRI and Bayesian network modeling study. *NeuroImage* **56** 1035–1042.

LI, Z., ROSEN, O., FERRARELLI, F. and KRAFTY, R. T. (2021). Adaptive Bayesian spectral analysis of high-dimensional nonstationary time series. *J. Comput. Graph. Statist.* 1–14. https://doi.org/10.1080/10618600.2020.1868305

LI, Z., YUE, Y. R. and BRUCE, S. A. (2024). ANOPOW for replicated nonstationary time series in experiments. *Ann. Appl. Stat.* **18** 328–349. https://doi.org/10.1214/23-AOAS1791

LORD, C., RUTTER, M. and COULTER, A. L. (1994). Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. Autism Dev. Disord.* **24** 659–685.

LORD, C., RUTTER, M., GOODE, S., HEEMSBERGEN, J., JORDAN, H., MAWHOOD, L. and SCHOPLER, E. (1989). Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. *J. Autism Dev. Disord.* **19** 185–212.

MANGOLD, S. and DAS, J. (2023). *Neuroanatomy, Cortical Primary Auditory Area*. StatPearls.

MARS, R., NEUBERT, F.-X., NOONAN, M., SALLET, J., TONI, I. and RUSHWORTH, M. (2012). On the relationship between the "default mode network" and the "social brain". *Front. Human Neurosci.* **6**. https://doi.org/10.3389/fnhum.2012.00189

MARTINEZ, J. G., BOHN, K. M., CARROLL, R. J. and MORRIS, J. S. (2013). A study of Mexican free-tailed bat chirp syllables: Bayesian functional mixed models for nonstationary acoustic time series. *J. Amer. Statist. Assoc.* **108** 514–526. PMID: 23997376. https://doi.org/10.1080/01621459.2013.793118

MONK, C. S., PELTIER, S. J., WIGGINS, J. L., WENG, S.-J., CARRASCO, M., RISI, S. and LORD, C. (2009). Abnormalities of intrinsic functional connectivity in autism spectrum disorders. *NeuroImage* **47** 764–772. https://doi.org/10.1016/j.neuroimage.2009.04.069

NEJAD, A., FOSSATI, P. and LEMOGNE, C. (2013). Self-referential processing, rumination, and cortical midline structures in major depression. *Front. Human Neurosci.* **7** 666. https://doi.org/10.3389/fnhum.2013.00666

PORT, R. G., GANDAL, M. J., ROBERTS, T. P. L., SIEGEL, S. J. and CARLSON, G. C. (2014). Convergence of circuit dysfunction in ASD: A common bridge between diverse genetic and environmental risk factors and common clinical electrophysiology. *Front. Cell. Neurosci.* **8**.

QIN, L., GUO, W. and LITT, B. (2009). A time-frequency functional model for locally stationary time series data. *J. Comput. Graph. Statist.* **18** 675–693. PMID: 20228961. MR2572632 https://doi.org/10.1198/jcgs.2009.06109

RAICHLE, M. E. (2010). Two views of brain function. *Trends Cogn. Sci.* **14**.

SALVADOR, R., SUCKLING, J., COLEMAN, M. R., PICKARD, J. D., MENON, D. and BULLMORE, E. (2005). Neurophysiological architecture of functional magnetic resonance images of human brain. *Cereb. Cortex* **15** 1332–1342. https://doi.org/10.1093/cercor/bhi016

SHEN, W., TU, Y., GOLLUB, R. L., ORTIZ, A., NAPADOW, V., YU, S., WILSON, G., PARK, J., LANG, C. et al. (2019). Visual network alterations in brain functional connectivity in chronic low back pain: A resting state functional connectivity and machine learning study. *NeuroImage Clin.* **22** 101775. https://doi.org/10.1016/j.nicl.2019.101775

SHUMWAY, R. H. and STOFFER, D. S. (2017). *Time Series Analysis and Its Applications*, 4th ed. *Springer Texts in Statistics*. Springer, Cham. MR3642322 https://doi.org/10.1007/978-3-319-52452-8

TERRY, D. P., SABATINELLI, D., PUENTE, A. N., LAZAR, N. A. and MILLER, L. S. (2015). A meta-analysis of fMRI activation differences during episodic memory in Alzheimer's disease and mild cognitive impairment. *J. Neuroimaging* **25** 849–860. https://doi.org/10.1111/jon.12266

TING, C.-M., OMBAO, H., SAMDIN, S. B. and SALLEH, S.-H. (2018). Estimating dynamic connectivity states in fMRI using regime-switching factor models. *IEEE Trans. Med. Imag.* **37** 1011–1023. https://doi.org/10.1109/TMI.2017.2780185

TOGO, K. and IWASAKI, M. (2013). Group comparisons involving zero-inflated count data in clinical trials. *Jpn. J. Biom.* **34** 53–66.

VALDÉS-SOSA, P. A., SÁNCHEZ-BORNOT, J. M., LAGE-CASTELLANOS, A., VEGA-HERNÁNDEZ, M., BOSCH-BAYARD, J., MELIE-GARCÍA, L. and CANALES-RODRÍGUEZ, E. (2005). Estimating brain functional connectivity with sparse multivariate autoregression. *Philos. Trans. R. Soc. B, Biol. Sci.* **360** 969–981. https://doi.org/10.1098/rstb.2005.1654

WANG, J., BARSTEIN, J., ETHRIDGE, L., MOSCONI, M., TAKARAE, Y. and SWEENEY, J. (2013). Resting state EEG abnormalities in autism spectrum disorders. *J. Neurodevelopmental Disord.* **5**.

WANG, L., ALDIRAWI, H. and YANG, J. (2020). Identifying zero-inflated distributions with a new R package iZID. *Commun. Inf. Syst.* **20** 23–44.

WANG, Q., LI, H.-Y., LI, Y.-D., WANG, R. and DENG, W. (2021). Resting-state abnormalities in functional connectivity of the default mode network in autism spectrum disorder: A meta-analysis. *Brain Imaging Behav.* **15** 2583–2592. https://doi.org/10.1007/s11682-021-00460-5

WEBER, C. F., KEBETS, V., BENKARIM, O., LARIVIÈRE, S., WANG, Y., NGO, A., JIANG, H., CHAI, X., PARK, B. Y. et al. (2024). Contracted functional connectivity profiles in autism. *Mol. Autism* **15** 38. https://doi.org/10.1186/s13229-024-00616-2

XIANG, Z., SUNDARARAJAN, R. and OMBAO, H. (2026). Supplement to "High-dimensional locally stationary models for identifying neuroimaging biomarkers in Autism Spectrum Disorder." https://doi.org/10.1214/26-AOAS2143SUPPA, https://doi.org/10.1214/26-AOAS2143SUPPB

YAESOUBI, M., ALLEN, E. A., MILLER, R. L. and CALHOUN, V. D. (2015). Dynamic coherence analysis of resting fMRI data to jointly capture state-based phase, frequency, and time-domain information. *NeuroImage* **120** 133–142. https://doi.org/10.1016/j.neuroimage.2015.07.002

YE, A., GATES, K. M., HENRY, T. R. and LUO, L. (2021). Path and directionality discovery in individual dynamic models: A regularized unified structural equation modeling approach for hybrid vector autoregression. *Psychometrika* **86** 404–9441. https://doi.org/10.1007/s11336-021-09753-6

YE, J., LI, Y., LAZAR, N. A., SCHAEFFER, D. J. and MCDOWELL, J. E. (2016). Finding common task-related regions in fMRI data from multiple subjects by periodogram clustering and clustering ensemble. *Stat. Med.* **35** 2635–2651. https://doi.org/10.1002/sim.6906

ZHU, Y. and CRIBBEN, I. (2018). Sparse graphical models for functional connectivity networks: Best methods and the autocorrelation issue. *Brain Connect*. **8** 139–165. PMID: 29634321. https://doi.org/10.1089/brain.2017.0511

ZÜRCHER, N. R., BHANOT, A., MCDOUGLE, C. J. and HOOKER, J. M. (2015). A systematic review of molecular imaging (PET and SPECT) in autism spectrum disorder: Current state and future research opportunities. *Neurosci. Biobehav. Rev.* **52** 56–73. https://doi.org/10.1016/j.neubiorev.2015.02.002

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1477–1495
https://doi.org/10.1214/26-AOAS2157

COVARIANCE REGRESSION WITH HIGH-DIMENSIONAL PREDICTORS: AN APPLICATION TO LINK BRAIN STRUCTURAL AND FUNCTIONAL CONNECTIVITY

BY YUHENG HE[1,a], CHANGLIANG ZOU[1,b] AND YI ZHAO[2,c]

[1]*School of Statistics and Data Science, Nankai University,* [a]*yuheng_he@foxmail.com,* [b]*nk.chlzou@gmail.com*
[2]*Department of Biostatistics and Health Data Science, Indiana University School of Medicine,* [c]*yz125@iu.edu*

In the high-dimensional landscape, addressing the challenges of covariance regression with high-dimensional predictors has posed difficulties for conventional methodologies. This paper addresses these hurdles by presenting a novel approach for high-dimensional inference with covariance matrix outcomes. The proposed methodology is demonstrated through its application in identifying patterns of brain co-activation observed in functional magnetic resonance imaging (fMRI) experiments and in revealing the predictive role of brain structural connectivity mapped through diffusion tensor imaging (DTI). In the pursuit of dependable statistical inference, we introduce an integrative approach based on penalized estimation. This approach combines data splitting, variable selection, aggregation of low-dimensional estimators, and robust variance estimation. It enables the construction of reliable confidence intervals for covariate coefficients, supported by theoretical confidence levels under specified conditions, where asymptotic distributions are provided. Through various simulation studies, the proposed approach performs well for covariance regression in the presence of high-dimensional predictors. This innovative approach is applied to the Lifespan Human Connectome Project (HCP) Aging Study. Brain networks and corresponding regions are identified, where regional DTI metrics predict within network resting-state functional connectivity. The findings are in line with established knowledge of the human brain.

REFERENCES

BARBER, R. F. and CANDÈS, E. J. (2019). A knockoff filter for high-dimensional selective inference. *Ann. Statist.* **47** 2504–2537. MR3988764 https://doi.org/10.1214/18-AOS1755

BASSETT, D. S. and BULLMORE, E. T. (2017). Small-world brain networks revisited. *Neuroscientist* **23** 499–516.

BASSETT, D. S. and SPORNS, O. (2017). Netw. Neurosci. *Nat. Neurosci.* **20** 353–364.

BOOKHEIMER, S. Y., SALAT, D. H., TERPSTRA, M., ANCES, B. M., BARCH, D. M., BUCKNER, R. L., BURGESS, G. C., CURTISS, S. W., DIAZ-SANTOS, M. et al. (2019). The lifespan human connectome project in aging: An overview. *NeuroImage* **185** 335–348.

BRESSLER, S. L. and MENON, V. (2010). Large-scale brain networks in cognition: Emerging methods and principles. *Trends Cogn. Sci.* **14** 277–290.

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

BULLMORE, E. and SPORNS, O. (2012). The economy of brain network organization. *Nat. Rev. Neurosci.* **13** 336.

CAI, J.-F., DONG, B., OSHER, S. and SHEN, Z. (2012). Image restoration: Total variation, wavelet frames, and beyond. *J. Amer. Math. Soc.* **25** 1033–1089. MR2947945 https://doi.org/10.1090/S0894-0347-2012-00740-1

COHEN, M. X. (2014). *Analyzing Neural Time Series Data*: *Theory and Practice*. MIT Press, Cambridge.

DAI, C., LIN, B., XING, X. and LIU, J. S. (2023). False discovery rate control via data splitting. *J. Amer. Statist. Assoc.* **118** 2503–2520. MR4681600 https://doi.org/10.1080/01621459.2022.2060113

DECO, G., PONCE-ALVAREZ, A., MANTINI, D., ROMANI, G. L., HAGMANN, P. and CORBETTA, M. (2013). Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. *J. Neurosci.* **33** 11239–11252.

Key words and phrases. Covariance regression, confidence interval, high-dimensional predictors, penalized estimation, variance estimation.

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

DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464

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

EFRON, B. (2014). Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.* **109** 991–1007. MR3265671 https://doi.org/10.1080/01621459.2013.823775

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

FEI, Z. and LI, Y. (2021). Estimation and inference for high dimensional generalized linear models: A splitting and smoothing approach. *J. Mach. Learn. Res.* **22** Paper No. 58. MR4253751

FLURY, B. N. (1986). Asymptotic theory for common principal component analysis. *Ann. Statist.* **14** 418–430. MR0840506 https://doi.org/10.1214/aos/1176349930

FORNITO, A., ZALESKY, A. and BREAKSPEAR, M. (2015). The connectomics of brain disorders. *Nat. Rev. Neurosci.* **16** 159.

FRISTON, K. J. (1994). Functional and effective connectivity in neuroimaging: A synthesis. *Hum. Brain Mapp.* **2** 56–78.

GLASSER, M. F., SOTIROPOULOS, S. N., WILSON, J. A., COALSON, T. S., FISCHL, B., ANDERSSON, J. L., XU, J., JBABDI, S., WEBSTER, M. et al. (2013). The minimal preprocessing pipelines for the human connectome project. *NeuroImage* **80** 105–124.

GOLLO, L. L., ROBERTS, J. A., CROPLEY, V. L., DI BIASE, M. A., PANTELIS, C., ZALESKY, A. and BREAKSPEAR, M. (2018). Fragility and volatility of structural hubs in the human connectome. *Nat. Neurosci.* **21** 1107–1116.

GROSENICK, L., KLINGENBERG, B., KATOVICH, K., KNUTSON, B. and TAYLOR, J. E. (2013). Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage* **72** 304–321.

HAGMANN, P., SPORNS, O., MADAN, N., CAMMOUN, L., PIENAAR, R., WEDEEN, V. J., MEULI, R., THIRAN, J.-P. and GRANT, P. (2010). White matter maturation reshapes structural connectivity in the late developing human brain. *Proc. Natl. Acad. Sci. USA* **107** 19067–19072.

HALL, Z., CHIEN, B., ZHAO, Y., RISACHER, S. L., SAYKIN, A. J., WU, Y.-C. and WEN, Q. (2022). Tau deposition and structural connectivity demonstrate differential association patterns with neurocognitive tests. *Brain Imaging Behav.* **16** 702–714.

HE, Y., ZOU, C. and ZHAO, Y. (2026). Supplement to "Covariance regression with high-dimensional predictors: an spplication to link brain structural and functional connectivity." https://doi.org/10.1214/26-AOAS2157SUPPA, https://doi.org/10.1214/26-AOAS2157SUPPB

HEBB, D. O. (2005). *The Organization of Behavior*: *A Neuropsychological Theory*. Psychology Press.

HERMUNDSTAD, A. M., BASSETT, D. S., BROWN, K. S., AMINOFF, E. M., CLEWETT, D., FREEMAN, S., FRITHSEN, A., JOHNSON, A., TIPPER, C. M. et al. (2013). Structural foundations of resting-state and task-based functional connectivity in the human brain. *Proc. Natl. Acad. Sci. USA* **110** 6169–6174.

HONEY, C., SPORNS, O., CAMMOUN, L., GIGANDET, X., THIRAN, J.-P., MEULI, R. and HAGMANN, P. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proc. Natl. Acad. Sci. USA* **106** 2035–2040.

HUANG, C., SUN, X., XIONG, J. and YAO, Y. (2016). Split LBI: An iterative regularization path with structural sparsity. *Adv. Neural Inf. Process. Syst.* **29**.

HUANG, C., SUN, X., XIONG, J. and YAO, Y. (2020). Boosting with structural sparsity: A differential inclusion approach. *Appl. Comput. Harmon. Anal.* **48** 1–45. MR4016983 https://doi.org/10.1016/j.acha.2017.12.004

HUANG, S., LI, J., YE, J., FLEISHER, A., CHEN, K., WU, T., REIMAN, E. and THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2012). A sparse structure learning algorithm for Gaussian Bayesian network identification from high-dimensional data. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 1328–1342.

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

JBABDI, S. and JOHANSEN-BERG, H. (2011). Tractography: Where do we go from here? *Brain Connect.* **1** 169–183.

JONES, D. K. (2011). *Diffusion MRI*: *Theory*, *Methods*, *and Application*. Oxford Univ. Press, New York.

KIM, R. and ZHANG, J. (2024). High-dimensional covariance regression with application to co-expression QTL detection. arXiv Preprint. Available at arXiv:2404.02093.

KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. MR2505584 https://doi.org/10.1137/070690274

KRZANOWSKI, W. (1984). Principal component analysis in the presence of group structure. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **33** 164–168.

LAIRD, A. R., EICKHOFF, S. B., LI, K., ROBIN, D. A., GLAHN, D. C. and FOX, P. T. (2009). Investigating the functional heterogeneity of the default mode network using coordinate-based meta-analytic modeling. *J. Neurosci.* **29** 14496–14505.

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371

LEECH, R., KAMOURIEH, S., BECKMANN, C. F. and SHARP, D. J. (2011). Fractionating the default mode network: Distinct contributions of the ventral and dorsal posterior cingulate cortex to cognitive control. *J. Neurosci.* **31** 3217–3224.

LE BIHAN, D. (2003). Looking into the functional architecture of the brain with diffusion MRI. *Nat. Rev. Neurosci.* **4** 469–480.

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

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523 https://doi.org/10.1111/j.1467-9868.2010.00740.x

MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. MR1803168 https://doi.org/10.2307/2669386

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133 https://doi.org/10.1214/12-STS400

PRETI, M. G. and VAN DE VILLE, D. (2019). Decoupling of brain function from structure reveals regional behavioral specialization in humans. *Nat. Commun.* **10** 4747.

RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D, Nonlinear Phenom.* **60** 259–268.

SMITH, S. M., JENKINSON, M., WOOLRICH, M. W., BECKMANN, C. F., BEHRENS, T. E., JOHANSEN-BERG, H., BANNISTER, P. R., DE LUCA, M., DROBNJAK, I. et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23** S208–S219.

SONG, H., DAI, R., RASKUTTI, G. and BARBER, R. F. (2020). Convex and non-convex approaches for statistical inference with class-conditional noisy labels. *J. Mach. Learn. Res.* **21** Paper No. 168. MR4209454

SPORNS, O. (2007). Brain connect.. *Scholarpedia* **2** 4695.

SUÁREZ, L. E., MARKELLO, R. D., BETZEL, R. F. and MISIC, B. (2020). Linking structure and function in macroscale brain networks. *Trends Cogn. Sci.* **24** 302–315.

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

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

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

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

VAN DEN HEUVEL, M. P., SPORNS, O., COLLIN, G., SCHEEWE, T., MANDL, R. C., CAHN, W., GOÑI, J., POL, H. E. H. and KAHN, R. S. (2013). Abnormal rich club organization and functional brain dynamics in schizophrenia. *JAMA Psychiatr.* **70** 783–792.

WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689 https://doi.org/10.1214/08-AOS646

WEDEEN, V. J., HAGMANN, P., TSENG, W.-Y. I., REESE, T. G. and WEISSKOFF, R. M. (2005). Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging. *Magn. Reson. Med.* **54** 1377–1386.

XIA, C. H., MA, Z., CIRIC, R., GU, S., BETZEL, R. F., KACZKURKIN, A. N., CALKINS, M. E., COOK, P. A., DE LA GARZA, A. G. et al. (2018). Linked dimensions of psychopathology and connectivity in functional brain networks. *Nat. Commun.* **9** 3003.

XIA, L., NAN, B. and LI, Y. (2023). Debiased lasso for generalized linear models with a diverging number of covariates. *Biometrics* **79** 344–357. MR4572526 https://doi.org/10.1111/biom.13587

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

ZHANG, Z., LIAO, W., CHEN, H., MANTINI, D., DING, J.-R., XU, Q., WANG, Z., YUAN, C., CHEN, G. et al. (2011). Altered functional–structural coupling of large-scale brain networks in idiopathic generalized epilepsy. *Brain* **134** 2912–2928.

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

ZHAO, Y., WANG, B., MOSTOFSKY, S. H., CAFFO, B. S. and LUO, X. (2021). Covariate assisted principal regression for covariance matrix outcomes. *Biostatistics* **22** 629–645. MR4287172 https://doi.org/10.1093/biostatistics/kxz057

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1496–1515
https://doi.org/10.1214/26-AOAS2140

BSNMANI: BAYESIAN SCALAR-ON-NETWORK REGRESSION WITH MANIFOLD LEARNING

BY YIJUN LI[1,a], KI SUENG CHOI[2,c], BOADIE W. DUNLOP[3,e], W. EDWARD CRAIGHEAD[3,f], HELEN S. MAYBERG[2,d], LANA GARMIRE[4,g], YING GUO[5,h] AND JIAN KANG[1,b]

[1]*Department of Biostatistics, University of Michigan,* [a]*liyijun@umich.edu,* [b]*jiankang@umich.edu*
[2]*Center of Advanced Circuit Therapeutics, Icahn School of Medicine at Mount Sinai,* [c]*kisueng.choi@mssm.edu,* [d]*helen.mayberg@mssm.edu*
[3]*Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine,* [e]*bdunlop@emory.edu,* [f]*ecraigh@emory.edu*
[4]*Department of Computational Medicine and Bioinformatics, University of Michigan,* [g]*lgarmire@med.umich.edu*
[5]*Department of Biostatistics and Bioinformatics, Emory University,* [h]*yguo2@emory.edu*

Brain connectivity analysis is crucial for understanding brain structure and neurological function, shedding light on the mechanisms of mental illness. To study the association between individual brain connectivity networks and the clinical characteristics, we develop BSNMani: a Bayesian scalar-on-network regression model with manifold learning. BSNMani comprises two components: the network manifold learning model for brain connectivity networks, which extracts shared connectivity structures and subject-specific network features, and the joint predictive model for clinical outcomes, which studies the association between clinical phenotypes and subject-specific network features while adjusting for potential confounding covariates. For posterior computation, we develop a novel two-stage hybrid algorithm combining Metropolis-Adjusted Langevin Algorithm (MALA) and Gibbs sampling. Our method is not only able to extract meaningful subnetwork features that reveal shared connectivity patterns but can also reveal their association with clinical phenotypes, further enabling clinical outcome prediction. We demonstrate our method through simulations and through its application to real resting-state fMRI data from a study focusing on Major Depressive Disorder (MDD). Our approach sheds light on the intricate interplay between brain connectivity and clinical features, offering insights that can contribute to our understanding of psychiatric and neurological disorders as well as mental health.

REFERENCES

AMICO, E., MARINAZZO, D., DI PERRI, C., HEINE, L., ANNEN, J., MARTIAL, C., DZEMIDZIC, M., KIRSCH, M., BONHOMME, V. et al. (2017). Mapping the functional connectome traits of levels of consciousness. *NeuroImage* **148** 201–211.

BRZYSKI, D., HU, X., GOÑI, J., ANCES, B., RANDOLPH, T. W. and HAREZLAK, J. (2023). Matrix-variate regression for sparse, low-rank estimation of brain connectivities associated with a clinical outcome. *IEEE Trans. Biomed. Eng.* **71** 1378–1390.

CASH, R. F., COCCHI, L., LV, J., FITZGERALD, P. B. and ZALESKY, A. (2021). Functional magnetic resonance imaging–guided personalization of transcranial magnetic stimulation treatment for depression. *JAMA Psychiatr.* **78** 337–339.

CHEN, Y., WANG, X., KONG, L. and ZHU, H. (2016). Local region sparse learning for image-on-scalar regression. arXiv preprint. Available at arXiv:1605.08501.

COHEN, S. E., ZANTVOORD, J. B., WEZENBERG, B. N., BOCKTING, C. L. and VAN WINGEN, G. A. (2021). Magnetic resonance imaging for individual prediction of treatment response in major depressive disorder: A systematic review and meta-analysis. *Transl. Psychiatry* **11** 168.

DUNLOP, B. W., BINDER, E. B., CUBELLS, J. F., GOODMAN, M. M., KELLEY, M. E., KINKEAD, B., KUTNER, M., NEMEROFF, C. B., NEWPORT, D. J. et al. (2012). Predictors of remission in depression to individual and combined treatments (PReDICT): Study protocol for a randomized controlled trial. *Trials* **13** 1–18.

Key words and phrases. Neuroimaging, brain connectivity, manifold learning, major depressive disorder.

DUNLOP, B. W., CHA, J., CHOI, K. S., RAJENDRA, J. K., NEMEROFF, C. B., CRAIGHEAD, W. E. and MAYBERG, H. S. (2023). Shared and unique changes in brain connectivity among depressed patients after remission with pharmacotherapy versus psychotherapy. *Amer. J. Psychiatry* **180** 218–229.

DUNLOP, B. W., COLE, S. P., NEMEROFF, C. B., MAYBERG, H. S. and CRAIGHEAD, W. E. (2018). Differential change on depressive symptom factors with antidepressant medication and cognitive behavior therapy for major depressive disorder. *J. Affective Disorders* **229** 111–119.

DURANTE, D., DUNSON, D. B. and VOGELSTEIN, J. T. (2017b). Nonparametric Bayes modeling of populations of networks. *J. Amer. Statist. Assoc.* **112** 1516–1530.

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

GORDON, E. M., LAUMANN, T. O., ADEYEMO, B., HUCKINS, J. F., KELLEY, W. M. and PETERSEN, S. E. (2016). Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cereb. Cortex* **26** 288–303.

GUHA, S. and GUHANIYOGI, R. (2021). Bayesian generalized sparse symmetric tensor-on-vector regression. *Technometrics* **63** 160–170.

GUHA, S. and GUHANIYOGI, R. (2024). Covariate-dependent clustering of undirected networks with brain-imaging data. *Technometrics* **66** 422–437.

GUHA, S. and RODRIGUEZ, A. (2023b). High-dimensional Bayesian network classification with network global-local shrinkage priors. *Bayesian Anal.* **18** 1131–1160.

HAMILTON, M. (1960). A rating scale for depression. *J. Neurol. Neurosurg. Psychiatry* **23** 56.

JAUCH, M., HOFF, P. D. and DUNSON, D. B. (2021). Monte Carlo simulation on the Stiefel manifold via polar expansion. *J. Comput. Graph. Statist.* **30** 622–631.

JIANG, B., PETKOVA, E., TARPEY, T. and OGDEN, R. T. (2020b). A Bayesian approach to joint modeling of matrix-valued imaging data and treatment outcome with applications to depression studies. *Biometrics* **76** 87–97.

JU, X., PARK, H. G. and TARPEY, T. (2025b). Bayesian scalar-on-network regression with applications to brain functional connectivity. *Biometrics* **81** ujaf023.

LI, Y., CHOI, K. S, DUNLOP, B. W, CRAIGHEAD, W. E, MAYBERG, H. S, GARMIRE, L., GUO, Y. and KANG, J. (2026). Supplement to "BSNMani: Bayesian scalar-on-network regression with manifold learning." https://doi.org/10.1214/26-AOAS2140SUPP

MA, X., KUNDU, S. and STEVENS, J. (2022b). Semi-parametric Bayes regression with network-valued covariates. *Mach. Learn.* **111** 3733–3767.

MAC GIOLLABHUI, N., MISCHOULON, D., DUNLOP, B. W., KINKEAD, B., SCHETTLER, P. J., LIU, R. T., OKEREKE, O. I., LAMON-FAVA, S., FAVA, M. et al. (2023). Individuals with depression exhibiting a pro-inflammatory phenotype receiving omega-3 polyunsaturated fatty acids experience improved motivation-related cognitive function: Preliminary results from a randomized controlled trial. *Brain Behav. Immun., Health* **32** 100666.

MORRIS, E. L., HE, K. and KANG, J. (2022b). Scalar on network regression via boosting. *Ann. Appl. Stat.* **16** 2755.

POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M. et al. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.

REHM, J. and SHIELD, K. D. (2019). Global burden of disease and the impact of mental and addictive disorders. *Curr. Psychiatry Rep.* **21** 10.

SCHAEFER, A., KONG, R., GORDON, E. M., LAUMANN, T. O., ZUO, X.-N., HOLMES, A. J., EICKHOFF, S. B. and YEO, B. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* **28** 3095–3114.

SUN, W. W. and LI, L. (2017b). Store: Sparse tensor response regression and neuroimaging analysis. *J. Mach. Learn. Res.* **18** 4908–4944.

TIAN, Y., MARGULIES, D. S., BREAKSPEAR, M. and ZALESKY, A. (2020). Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nat. Neurosci.* **23** 1421–1432.

VAROQUAUX, G., BARONNET, F., KLEINSCHMIDT, A., FILLARD, P. and THIRION, B. (2010). Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 200–208. Springer, Berlin.

WANG, L., DURANTE, D., JUNG, R. E. and DUNSON, D. B. (2017). Bayesian network–response regression. *Bioinformatics* **33** 1859–1866.

WANG, L., ZHANG, Z. and DUNSON, D. (2019b). Symmetric bilinear regression for signal subgraph estimation. *IEEE Trans. Signal Process.* **67** 1929–1940.

WANG, W., ZHANG, X. and LI, L. (2019b). Common reducing subspace model and network alternation analysis. *Biometrics* **75** 1109–1120.

WANG, Y. and GUO, Y. (2023a). LOCUS: A regularized blind source separation method with low-rank structure for investigating brain connectivity. *Ann. Appl. Stat.* **17** 1307–1332.

WELLING, M. and TEH, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the* 28*th International Conference on Machine Learning* (*ICML*-11) 681–688. Citeseer.

ZHANG, D., LI, L., SRIPADA, C. and KANG, J. (2023). Image response regression via deep neural networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 1589–1614.

ZHU, H., FAN, J. and KONG, L. (2014b). Spatially varying coefficient model for neuroimaging data with jump discontinuities. *J. Amer. Statist. Assoc.* **109** 1084–1098.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1516–1540
https://doi.org/10.1214/26-AOAS2169

TOPOLOGICAL INFERENCE ON BRAIN NETWORKS WITH APPLICATION TO LESION SYMPTOM MAPPING

BY YUAN WANG[1,a], JIAN YIN[2,b], NICHOLAS RICCARDI[3,c], DIRK-BART DEN OUDEN[3,d], JULIUS FRIDRIKSSON[3,e] AND RUTVIK H. DESAI[4,f]

[1]*Department of Epidemiology and Biostatistics, University of South Carolina,* [a]*wang578@mailbox.sc.edu*
[2]*Department of Biostatistics, City University of Hong Kong,* [b]*jian.yin@my.cityu.edu.hk*
[3]*Department of Communication Science and Disorders, University of South Carolina,* [c]*riccardn@email.sc.edu,* [d]*ouden@mailbox.sc.edu,* [e]*fridriks@mailbox.sc.edu*
[4]*Department of Psychology, University of South Carolina,* [f]*rutvik@sc.edu*

Persistent homology (PH) characterizes the shape of brain networks through persistence features. Group comparison of persistence features from brain networks can be challenging, as they are inherently heterogeneous. A recent scale-space representation of persistence diagram (PD) through heat diffusion reparameterizes using a finite number of Fourier coefficients with respect to the Laplace–Beltrami (LB) eigenfunction expansion of the domain, thus providing a powerful vectorized algebraic representation for group comparisons of PDs. In this study, we advance a transposition-based permutation test for comparing multiple groups of PDs using their heat-diffusion estimates of the PDs. We evaluate the empirical performance of the spectral transposition test in capturing within- and between-group similarity and dissimilarity under statistical variation in topological noise and cycle location. In application, we introduce a *topological lesion symptom mapping* (TLSM) method based on the proposed topological inference framework. The method is applied to resting-state functional brain networks from individuals with post-stroke aphasia to identify characteristic cycles associated with varying degrees of speech-language impairment, as measured by behavioral test scores.

REFERENCES

ADAMS, H., EMERSON, T., KIRBY, M., NEVILLE, R., PETERSON, C., SHIPMAN, P., CHEPUSHTANOVA, S., HANSON, E., MOTTA, F. et al. (2017). Persistence images: A stable vector representation of persistent homology. *J. Mach. Learn. Res.* **18** 218–252. MR3625712

ALDOUS, D. (1983). Random walks on finite groups and rapidly mixing Markov chains. In *Seminar on Probability XVII* 1981/82 243–297. Springer, Berlin. MR0770418 https://doi.org/10.1007/BFb0068322

ALDOUS, D. and DIACONIS, P. (1986). Shuffling cards and stopping times. *Amer. Math. Monthly* **93** 333–348. MR0841111 https://doi.org/10.2307/2323590

ANDERSON, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26** 32–46.

ANDERSON, M. J. (2017). Permutational Multivariate Analysis of Variance (PERMANOVA). Wiley StatsRef: Statistics Reference Online.

ASHBURNER, J. and FRISTON, K. (2000). Voxel-based morphometry—the methods. *NeuroImage* **11** 805–821.

BASSETT, D. S. (2006). Adaptive reconfiguration of fractal small-world human brain functional networks. *Proc. Natl. Acad. Sci. USA* **203** 19518–23.

BATES, E., WILSON, S. M., SAYGIN, A. P. et al. (2003). Voxel-based lesion–symptom mapping. *Nat. Neurosci.* **6** 448–450.

BIEN, J. and TIBSHIRANI, R. (2011). Hierarchical clustering with prototypes via minimax linkage. *J. Amer. Statist. Assoc.* **106** 1075–1084. MR2894765 https://doi.org/10.1198/jasa.2011.tm10183

BUBENIK, P. (2015). Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **16** 77–102. MR3317230

Key words and phrases. Topological data analysis, persistent homology, permutation test, brain network, lesion symptom mapping.

CARRIÉRE, M., OUDOT, S. and OVSJANIKOV, M. (2015). Stable topological signatures for points on 3D shapes. In *Eurographics Symposium on Geometry Processing*, *Vol*. 34.

CHEN, Y.-C., WANG, D., RINALDO, A. and WASSERMAN, L. (2015). Statistical analysis of persistence intensity functions. Available at arXiv:1510.02502v1.

CHUNG, M. K., BUBENIK, P. and KIM, P. (2009). Persistence diagrams of cortical surface data. In *Proceedings of Information Processing in Medical Imaging* (*IPMI*) 386–397.

CHUNG, M. K., DALTON, K. M., SHEN, L., EVANS, A. C. and DAVIDSON, R. J. (2007). Weighted Fourier representation and its application to quantifying the amount of gray matter. *IEEE Trans*. *Med*. *Imag*. **26** 566–581.

CHUNG, M. K., HANSON, J. L., YE, J., DAVIDSON, R. J. and POLLAK, S. D. (2015). Persistent homology in sparse regression and its application to brain morphometry. *IEEE Trans*. *Med*. *Imag*. **34**. 1928–1939.

CHUNG, M. K., LEE, H., OMBAO, H. and SOLO, V. (2019a). Exact topological inference of the resting-state brain networks in twins. *Netw*. *Neurosci*.

CHUNG, M. K., SCHAEFER, S. M., VAN REEKUM, C. M., SCHMITZ, L. P., SUTTERER, M. and DAVIDSON, R. J. (2014). A unified kernel regression on manifolds detects aging-related changes in the amygdala and hippocampus. In *MICCAI*. *Lecture Notes in Computer Science* (*LNCS*) **8674** 789–796.

CHUNG, M. K., WANG, Y. and WU, G. (2018). Heat kernel smoothing in irregular image domains. In *International Conference of the IEEE Engineering in Medicine and Biology Society* (*EMBC*) 5101–5104.

CHUNG, M. K., XIE, L., HUANG, S. G., WANG, Y., YAN, J. and SHEN, L. (2019b). Rapid acceleration of the permutation test via transpositions. In *International Workshop on Connectomics in NeuroImaging*. *Lecture Notes in Computer Science* **11848** 42–53.

EDELSBRUNNER, H., LETSCHER, D. and ZOMORODIAN, A. (2002). Topological persistence and simplification *Discrete Comput*. *Geom*. 511–533. MR1949898 https://doi.org/10.1007/s00454-002-2885-2

FASY, B. T., LECCI, F., RINALDO, A., WASSERMAN, L., BALAKRISHNAN, S. and SINGH, A. (2014). Confidence sets for persistence diagrams. *Ann*. *Statist*. **42** 2301–2339. MR3269981 https://doi.org/10.1214/14-AOS1252

GLEICHGERRCHT, E., FRIDRIKSSON, J., RORDEN, C. and BONILHA, L. (2017). Connectome-based lesion-symptom mapping (CLSM): A novel approach to map neurological function. *NeuroImage Clin*. **16** 461–467.

HATCHER, A. (2001). *Algebraic Topology*. Cambridge Univ. Press, Cambridge. MR1867354

HEO, G., GAMBLE, J. and KIM, P. T. (2012). Topological analysis of variance and the maxillary complex. *J*. *Amer*. *Statist*. *Assoc*. **107** 477–492. MR2980059 https://doi.org/10.1080/01621459.2011.641430

HUANG, S. G., CHUNG, M. K., QIU, A. and ADNI (2021). Fast mesh data augmentation via Chebyshev polynomial of spectral filtering. *Neural Netw*. **143** 198–208.

IVANOVA, M. V., HERRON, T. J., DRONKERS, N. F. and BALDO, J. V. (2021). An empirical comparison of univariate versus multivariate methods for the analysis of brain–behavior mapping. *Hum*. *Brain Mapp*. **42** 1070–1101.

KARNATH, H.-O., SPERBER, C. and RORDEN, C. (2018). Mapping human brain lesions and their functional consequences. *NeuroImage* **165** 180–189.

KERTESZ, A. (2007). The Western Aphasia Battery—Revised.

KULKARNI, A. P., CHUNG, M. K., BENDLIN, B. B. and PRABHAKARAN, V. (2020). Investigating heritability across resting state brain networks via heat kernel smoothing on persistence diagrams. In *Proceedings of the IEEE International Symposium on Biomedical Imaging Workshops* 1–4.

LEE, H., CHUNG, M. K., KANG, H., KIM, B. N. and LEE, D. S. (2011). Computing the shape of brain networks using graph filtration and Gromov–Hausdorff metric. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (*MICCAI*) **14** 302–309.

LEE, H., CHUNG, M. K., KANG, H. and LEE, D. S. (2014). Hole detection in metabolic connectivity of Alzheimer's disease using k-Laplacian. In *International Conference on Medical Imaging Computing and Computer-Assisted Intervention* (*MICCAI*). *Lecture Notes in Computer Science* (*LNCS*) **8675** 297–304.

MATKOWSKI, J. and NIKODEM, K. (1994). An integral Jensen inequality for convex multifunctions. *Result*. *Math*. **26** 348–353. MR1300617 https://doi.org/10.1007/BF03323058

MILEYKO, Y., MUKHERJEE, S. and HARER, J. (2011). Probability measures on the space of persistence diagrams. *Inverse Probl*. **27** 124007, 22. MR2854323 https://doi.org/10.1088/0266-5611/27/12/124007

PACHAURI, D., HINRICHS, C., CHUNG, M. K., JOHNSON, S. C. and SINGH, V. (2011). Topology-based kernels with application to inference problems in Alzheimer's disease. *IEEE Trans*. *Med*. *Imag*. **30** 1760–1770.

REININGHAUS, J., HUBER, S., BAUER, U. and KWITT, R. (2015). A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 4741–4748.

RICCARDI, N., ZHAO, X., DEN OUDEN, D. B., FRIDRIKSSON, J., DESAI, R. H. and WANG, Y. (2024). Network-based statistics distinguish anomic and Broca's aphasia. *Brain Struct*. *Funct*. **229** 2237–2253.

RORDEN, C. and KARNATH, H.-O. (2004). Using human brain lesions to infer function: A relic from a past era in the fMRI age? *Nat*. *Rev*. *Neurosci*. **5** 813–819.

SIZEMORE, A. E., GIUSTI, C., KAHN, A., VETTEL, J. M., BETZEL, R. F. and BASSETT, D. S. (2018). Cliques and cavities in the human connectome. *J. Comput. Neurosci.* **44** 115–145. MR3746561 https://doi.org/10.1007/s10827-017-0672-6

SONGDECHAKRAIWUT, T. and CHUNG, M. K. (2023). Topological learning for brain networks. *Ann. Appl. Stat.* **17** 403–433. MR4539037 https://doi.org/10.1214/22-aoas1633

TSAO, C. W., ADAY, A. W., ALMARZOOQ, Z. I., ALONSO, A., BEATON, A. Z., BITTENCOURT, M. S., BOEHME, A. K., BUXTON, A. E., CARSON, A. P. et al. (2022). Heart disease and stroke statistics—2022 update: A report from the American Heart Association. *Circulation* **145** e153–e639.

WANG, Y., BEHROOZMAND, R., PHILLIP JOHNSON, L., BONILHA, L. and FRIDRIKSSON, J. (2021). Topological signal processing and inference of event-related potential response. *J. Neurosci. Methods* **363** 109324.

WANG, Y., CHUNG, M. K. and FRIDRIKSSON, J. (2022). Spectral permutation test on persistence diagrams. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*) 1461–1465.

WANG, Y., OMBAO, H. and CHUNG, M. K. (2018). Topological data analysis of single-trial electroencephalographic signals. *Ann. Appl. Stat.* **12** 1506–1534. MR3852686 https://doi.org/10.1214/17-AOAS1119

WANG, Y., OMBAO, H. and CHUNG, M. K. (2019). Statistical persistent homology of brain signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*) 1125–1129.

WANG, Y., YIN, J. and DESAI, R. H. (2023). Topological inference on brain networks across subtypes of post-stroke aphasia. arXiv. arXiv–2311.

WANG, Y., YIN, J., RICCARDI, N., DEN OUDEN, D.-B., FRIDRIKSSON, J. and DESAI, R. H. (2026). Supplement to "Topological inference on brain networks with application to lesion symptom mapping." https://doi.org/10.1214/26-AOAS2169SUPP

YI, J., YIN, J., GHOSAL, R., DEN OUDEN, D., FRIDRIKSSON, J., DESAI, R. H. and WANG, Y. (2025a). Topological clustering of aphasic brain networks. In *NeurIPS Workshop on New Perspectives in Graph Machine Learning*.

YI, J., YIN, J., GHOSAL, R., FRIDRIKSSON, J., DESAI, R. H. and WANG, Y. (2025b). Uncovering treatment response patterns with topological clustering of brain networks. In *IEEE International Conference on Big Data* (*BigData*) 1269–1277.

YOURGANOV, G., FRIDRIKSSON, J., RORDEN, C., GLEICHGERRCHT, E. and BONILHA, L. (2016). Multivariate connectome-based symptom mapping in post-stroke patients: Networks supporting language and speech. *J. Neurosci.* **36** 6668–6679.

YOURGANOV, G., FRIDRIKSSON, J., STARK, B. and RORDEN, C. (2018). Removal of artifacts from resting-state fMRI data in stroke. *NeuroImage Clin.* **17** 297–305.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1541–1561
https://doi.org/10.1214/26-AOAS2185

GRANGER CAUSALITY FOR MIXED TIME SERIES GENERALIZED LINEAR MODELS: A CASE STUDY ON MULTIMODAL BRAIN CONNECTIVITY

BY LUIZA S. C. PIANCASTELLI[1,a], WAGNER BARRETO-SOUZA[1,b], NORBERT J. FORTIN[2,3,c], KEILAND W. COOPER[2,3,d] AND HERNANDO OMBAO[4,5,e]

[1]*School of Mathematics and Statistics, University College Dublin,* [a]*luiza.piancastelli@ucd.ie,* [b]*wagner.barreto-souza@ucd.ie*
[2]*Center for the Neurobiology of Learning and Memory, University of California*
[3]*Department of Neurobiology and Behavior, University of California,* [c]*norbert.fortin@uci.edu,* [d]*kwcooper@uci.edu*
[4]*Statistics Program, King Abdullah University of Science and Technology*
[5]*Neuroscience-AI Laboratory, King Abdullah University of Science and Technology,* [e]*hernando.ombao@kaust.edu.sa*

This paper is motivated by neuroscience studies aimed at understanding causal or predictive interactions between nodes in a brain network using multimodal brain activity data. To assess Granger causality, we introduce a flexible framework through a general class of models that accommodate mixed types of data (binary, count, continuous and positive components) formulated in a generalized linear model (GLM) fashion. To conduct statistical inference for causality, we propose a Bayesian mixed time series model that incorporates spike-and-slab priors on selected parameters. This framework enables effective selection of causal ordering and offers robust uncertainty quantification. The proposed methods are then applied to brain activity data, including both spike train and local field potential (LFP) activity, recorded as animals (rats) performed a complex sequence memory task. The proposed methodology provides critical insights into the causal relationship between band-specific spectral power in the LFP and subsequent spiking activity. Specifically, power in the LFP beta band is predictive of spiking activity 300 milliseconds later, providing a novel analytical tool for this area of emerging interest in neuroscience and demonstrating its usefulness and flexibility in the study of causality in general.

REFERENCES

ALLEN, T. A., MORRIS, A. M., MATTFELD, A. T., STARK, C. E. L. and FORTIN, N. J. (2014). A sequence of events model of episodic memory shows parallels in rats and humans. *Hippocampus* **24** 1178–1188.

ALLEN, T. A., SALZ, D. M., MCKENZIE, S. and FORTIN, N. J. (2016). Nonspatial sequence coding in CA1 neurons. *J. Neurosci.* **5** 1547–1563.

BUZSÁKI, G. (2002). Theta oscillations in the hippocampus. *Neuron* **3** 325–340.

CHEN, C. W. S. and LEE, S. (2017). Bayesian causality test for integer-valued time series models with applications to climate and crime data. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **66** 797–814. MR3670418 https://doi.org/10.1111/rssc.12200

COLGIN, L. L. (2016). Rhythms of the hippocampal network. *Nat. Rev. Neurosci.* **4** 239–249.

FIECAS, M. and OMBAO, H. (2016). Modeling the evolution of dynamic brain processes during an associative learning experiment. *J. Amer. Statist. Assoc.* **111** 1440–1453. MR3601700 https://doi.org/10.1080/01621459.2016.1165683

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

GAO, X., SHEN, W., SHAHBABA, B., FORTIN, N. J. and OMBAO, H. (2020). Evolutionary state-space model and its application to time-frequency analysis of local field potentials. *Statist. Sinica* **30** 1561–1582. MR4257545 https://doi.org/10.5705/ss.202017.0420

GATTAS, S., ELIAS, G. A., JANECEK, J., YASSA, M. A. and FORTIN, N. J. (2022). Proximal CA1 20–40 Hz power dynamics reflect trial-specific information processing supporting nonspatial sequence memory. *eLife* **11** e55528.

Key words and phrases. Bayesian model, brain signals, causality, multimodal data, time series.

GONG, X., LI, W. and LIANG, H. (2019). Spike-field Granger causality for hybrid neural data analysis. *NeuroImage* **122** 809–822.

GRANADOS-GARCIA, G., FIECAS, M., SHAHBABA, B., FORTIN, N. J. and OMBAO, H. (2022). Modeling brain waves as a mixture of latent processes. *Comput. Statist. Data Anal.* **174** 107409.

GRANGER, C. W. J. (1969). Investigating causal relations by econometric models and cross spectral methods. *Econometrica* **37** 424–438.

GRANGER, C. W. J., ROBINS, R. P. and ENGLE, R. F. (1986). Wholesale and retail prices: Bivariate time-series modeling with forecastable error variances. In *Model Reliability* (D. A. Belsley and E. Kuh, eds.) 1–17. MIT Press, Cambridge, MA.

GUO, S., LING, S. and ZHU, K. (2014). Factor double autoregressive models with application to simultaneous causality testing. *J. Statist. Plann. Inference* **148** 82–94. MR3174149 https://doi.org/10.1016/j.jspi.2013.12.007

HONG, Y., LIU, Y. and WANG, S. (2009). Granger causality in risk and detection of extreme risk spillover between financial markets. *J. Econometrics* **150** 271–287. MR2535522 https://doi.org/10.1016/j.jeconom.2008.12.013

HU, M., LI, M., LI, W. and LIANG, H. (2016). Joint analysis of spikes and local field potentials using copula. *NeuroImage* **133** 457–467.

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

LEE, B.-S. (1992). Causal relations among stock returns, interest rates, real activity, and inflation. *J. Finance* **47** 1591–1603.

LEE, Y. and LEE, S. (2019). On causality test for time series of counts based on Poisson INGARCH models with application to crime and temperature data. *Comm. Statist. Simulation Comput.* **48** 1901–1911. MR3945963 https://doi.org/10.1080/03610918.2018.1429618

LI, D., CHAN, N. H. and PENG, L. (2014). Empirical likelihood test for causality of bivariate AR(1) processes. *Econometric Theory* **30** 357–371. MR3231495 https://doi.org/10.1017/S0266466613000339

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

OMBAO, H. and PINTO, M. (2024). Spectral dependence. *Econom. Stat.* **32** 122–159. MR4799608 https://doi.org/10.1016/j.ecosta.2022.10.005

PANINSKI, L., BROWN, E. N., IYENGAR, S. and KASS, R. E. (2009). Statistical models of spike trains. In *Stochastic Methods in Neuroscience* (C. Laing and G. Lord, eds.). Oxford Univ. Press, Oxford. MR2642703

PIANCASTELLI, L. S. C., BARRETO-SOUZA, W. and OMBAO, H. (2023). Flexible bivariate INGARCH process with a broad range of contemporaneous correlation. *J. Time Series Anal.* **44** 206–222. MR4562933 https://doi.org/10.1111/jtsa.12663

PIANCASTELLI, L. S., BARRETO-SOUZA, W., FORTIN, N. J., COOPER, K. W. and OMBAO, H. (2026). Supplement to "Granger causality for mixed time series generalized linear models: a case study on multimodal brain connectivity." https://doi.org/10.1214/26-AOAS2185SUPP

SCHULTHEISS, N. W., SCHLECHT, M., JAYACHANDRAN, M., BROOKS, D. R., MCGLOTHAN, J. L., GUILARTE, T. R. and ALLEN, T. A. (2020). Awake delta and theta-rhythmic hippocampal network modes during intermittent locomotor behaviors in the rat. *Behav. Neurosci.* **6** 529–546.

SETH, A. K., BARRETT, A. B. and BARNETT, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *J. Neurosci.* **35** 3293–3297.

SHAHBABA, B., LI, L., AGOSTINELLI, F., SARAF, M., COOPER, K. W., HAGHVERDIAN, D., ELIAS, G. A., BALDI, P. and FORTIN, N. J. (2022). Hippocampal ensembles represent sequential relationships among an extended sequence of nonspatial events. *Nat. Commun.* **13** 787.

SHOJAIE, A. and FOX, E. B. (2022). Granger causality: A review and recent advances. *Annu. Rev. Stat. Appl.* **9** 289–319. MR4394910 https://doi.org/10.1146/annurev-statistics-040120-010930

SHUMWAY, R. H. and STOFFER, D. S. (2017). *Time Series Analysis and Its Applications*: *With R Examples*, 4th ed. *Springer Texts in Statistics*. Springer, Cham. MR3642322 https://doi.org/10.1007/978-3-319-52452-8

SIMS, C. A. (1980). Macroeconomics and reality. *Econometrica* **18** 1–48.

TANK, A., LI, X., FOX, E. B. and SHOJAIE, A. (2021). The convex mixture distribution: Granger causality for categorical time series. *SIAM J. Math. Data Sci.* **3** 83–112. MR4205091 https://doi.org/10.1137/20M133097X

ZHOU, B., MOORMAN, D. E., BEHSETA, S., OMBAO, H. and SHAHBABA, B. (2016). A dynamic Bayesian model for characterizing cross-neuronal interactions during decision-making. *J. Amer. Statist. Assoc.* **111** 459–471. MR3538679 https://doi.org/10.1080/01621459.2015.1116988

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1562–1585
https://doi.org/10.1214/26-AOAS2148

EXPLAINABLE PARAMETER CALIBRATION VIA IMPORTANCE-DRIVEN SEQUENTIAL DESIGN WITH AN APPLICATION TO BUILDING ENERGY SYSTEMS

BY CHEOLJOON JEONG[1,a] AND EUNSHIN BYON[2,b]

[1]*Department of Industrial Engineering, Clemson University,* [a]*cjeong@clemson.edu*
[2]*Department of Industrial and Operations Engineering, University of Michigan,* [b]*ebyon@umich.edu*

Parameter calibration seeks to estimate unobservable parameters in a computer model by aligning field observations with computer model outputs. In the building energy sector, a physics-based computer model is developed to analyze building energy use, given various weather conditions and operational scenarios. To obtain accurate simulations, it is necessary to calibrate model parameters required for preconfiguration. Among various techniques, Bayesian optimization stands out for its potential but faces some challenges when handling a large number of parameters. A possible remedy is to focus selectively on influential parameters, thereby simplifying a complex, high-dimensional task into a more tractable, lower-dimensional endeavor. We develop a new method that ranks parameter importance to effectively enable stochastic dimension reduction by utilizing the multi-armed bandit approach. By accounting for unequal importance among parameters, our approach generates accurate surrogate models tailored to the reduced dimension and guides an efficient exploration of the parameter search space in the Bayesian optimization procedure. The numerical studies and building energy simulation case study demonstrate that the proposed approach achieves a significant improvement in both calibration accuracy and efficiency. Moreover, it capably identifies the influential parameters and explains their impact on the computer model, providing valuable insights into understanding the system's dynamics.

REFERENCES

BA, S., MYERS, W. R. and WANG, D. (2018). A sequential maximum projection design framework for computer experiments with inert factors. *Statist. Sinica* **28** 879–897. MR3791092

BADANIDIYURU, A., KLEINBERG, R. and SLIVKINS, A. (2018). Bandits with knapsacks. *J. ACM* **65** 1–55. MR3771540 https://doi.org/10.1145/3164539

BINOIS, M. and WYCOFF, N. (2022). A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization. *ACM Trans. Evol. Learn. Optim.* **2**.

BOOTH, A., CHOUDHARY, R. and SPIEGELHALTER, D. (2013). A hierarchical Bayesian framework for calibrating micro-level models with macro-level data. *J. Build. Perform. Simul.* **6** 293–318.

CALAMA-GONZÁLEZ, C. M., SYMONDS, P., PETROU, G., SUÁREZ, R. and LEÓN-RODRÍGUEZ, Á. L. (2021). Bayesian calibration of building energy models for uncertainty analysis through test cells monitoring. *Appl. Energy* **282** 116118.

CESA-BIANCHI, N. and LUGOSI, G. (2012). Combinatorial bandits. *J. Comput. System Sci.* **78** 1404–1422. MR2926141 https://doi.org/10.1016/j.jcss.2012.01.001

CHAKRABARTY, A., MADDALENA, E., QIAO, H. and LAUGHMAN, C. (2021). Scalable Bayesian optimization for model calibration: Case study on coupled building and HVAC dynamics. *Energy Build.* **253** 111460.

CHEN, X., LI, K., MI, Z. and WANG, F. (2024). An air conditioning control method based on non-measurement thermal comfort standard correction for demand response. *IEEE Trans. Ind. Appl.* **60** 4736–4748.

CHONG, A. and MENBERG, K. (2018). Guidelines for the Bayesian calibration of building energy models. *Energy Build.* **174** 527–547.

CHRISTENSEN, C., HOROWITZ, S. and U.S. DEPARTMENT OF ENERGY OFFICE OF ENERGY EFFICIENCY AND RENEWABLE ENERGY (2011). BEopt (Building Energy Optimization Tool) [SWR-05-41], Vers. 2.8.0.0.

Key words and phrases. Bayesian optimization, efficient global optimization, multi-armed bandit, sequential design.

COAKLEY, D., RAFTERY, P. and KEANE, M. (2014). A review of methods to match building energy simulation models to measured data. *Renew. Sustain. Energy Rev.* **37** 123–141.

DALLA VALLE, A., TOOSI, H. A., LEONFORTE, F., DEL PERO, C., LAVAGNA, M., CAMPIOLI, A. and ASTE, N. (2025). Measuring the impact of holistic energy retrofit strategies: Life cycle assessment aligned with level (s). *Energy Build.* 116357.

DINH, H. T., LEE, K.-H. and KIM, D. (2022). Supervised-learning-based hour-ahead demand response for a behavior-based home energy management system approximating MILP optimization. *Appl. Energy* **321** 119382.

DIOUANE, Y., PICHENY, V., LE RICHE, R. and SCOTTO DI PERROTOLO, A. (2023). TREGO: A trust-region framework for efficient global optimization. *J. Global Optim.* **86** 1–23. MR4578206 https://doi.org/10.1007/s10898-022-01245-w

ERIKSSON, D., PEARCE, M., GARDNER, J. R., TURNER, R. and POLOCZEK, M. (2019). Scalable global optimization via local Bayesian optimization. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

EZZAT, A. A., POURHABIB, A. and DING, Y. (2018). Sequential design for functional calibration of computer models. *Technometrics* **60** 286–296. MR3847166 https://doi.org/10.1080/00401706.2017.1377638

FENG, Y., DUAN, Q., CHEN, X., YAKKALI, S. S. and WANG, J. (2021). Space cooling energy usage prediction based on utility data for residential buildings using machine learning methods. *Appl. Energy* **291** 116814.

GONZÁLEZ-TORRES, M., PÉREZ-LOMBARD, L., CORONEL, J. F., MAESTRE, I. R. and YAN, D. (2022). A review on buildings energy information: Trends, end-uses, fuels and drivers. *Energy Rep.* **8** 626–637.

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

HIGDON, D., KENNEDY, M., CAVENDISH, J. C., CAFEO, J. A. and RYNE, R. D. (2004). Combining field data and computer simulations for calibration and prediction. *SIAM J. Sci. Comput.* **26** 448–466. MR2116355 https://doi.org/10.1137/S1064827503426693

HOMMA, T. and SALTELLI, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliab. Eng. Syst. Saf.* **52** 1–17.

IOOSS, B. and LEMAÎTRE, P. (2015). A review on global sensitivity analysis methods. In *Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications* 101–122.

JAIN, P., SHASHAANI, S. and BYON, E. (2023). Wake effect parameter calibration with large-scale field operational data using stochastic optimization. *Appl. Energy* **347** 121426.

JAIN, P., SHASHAANI, S. and BYON, E. (2025). Simulation model calibration with dynamic stratification and adaptive sampling. *J. Simul.* **19** 494–515.

JEONG, C. and BYON, E. (2024). Calibration of building energy computer models via bias-corrected iteratively reweighted least squares method. *Appl. Energy* **360** 122753.

JEONG, C. and BYON, E. (2026). Supplement to "Explainable parameter calibration via importance-driven sequential design with an application to building energy systems." https://doi.org/10.1214/26-AOAS2148SUPPA, https://doi.org/10.1214/26-AOAS2148SUPPB

JEONG, C., YUE, X. and CHUNG, S. (2026). Fed-joint: Joint modeling of nonlinear degradation signals and failure events for remaining useful life prediction using federated learning. *Reliab. Eng. Syst. Saf.* **267** 111833.

JEONG, C., ZIANG, X., BYON, E., BERAHAS, A. S. and CETIN, K. (2023). Multiblock parameter calibration in computer models. *INFORMS J. Data Sci.* **2** 116–137.

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

JOSEPH, V. R., GUL, E. and BA, S. (2015). Maximum projection designs for computer experiments. *Biometrika* **102** 371–380. MR3371010 https://doi.org/10.1093/biomet/asv002

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

KLEIN, S. A. et al. (2017). *TRNSYS* 18: *A Transient System Simulation Program Solar Energy Laboratory*. Univ. Wisconsin Press, Madison, USA.

KONTAR, R., SHI, N., YUE, X., CHUNG, S., BYON, E., CHOWDHURY, M., JIN, J., KONTAR, W., MASOUD, N. et al. (2021). The Internet of federated things (IoFT). *IEEE Access* **9** 156071–156113.

LI, C., GUPTA, S., RANA, S., NGUYEN, V., VENKATESH, S. and SHILTON, A. (2017). High-dimensional Bayesian optimization using dropout. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence* 2096–2102.

LI, Q., AUGENBROE, G. and BROWN, J. (2016). Assessment of linear emulators in lightweight Bayesian calibration of dynamic building energy models for parameter estimation and performance prediction. *Energy Build.* **124** 194–202.

LI, S., KO, Y. M. and BYON, E. (2021). Nonparametric importance sampling for wind turbine reliability analysis with stochastic computer models. *Ann. Appl. Stat.* **15** 1850–1871. MR4355079 https://doi.org/10.1214/21-aoas1490

LINKLETTER, C., BINGHAM, D., HENGARTNER, N., HIGDON, D. and YE, K. Q. (2006). Variable selection for Gaussian process models in computer experiments. *Technometrics* **48** 478–490. MR2328617 https://doi.org/10.1198/004017006000000228

LIU, B., YUE, X., BYON, E. and KONTAR, R. A. (2022). Parameter calibration in wake effect simulation model with stochastic gradient descent and stratified sampling. *Ann. Appl. Stat.* **16** 1795–1821. MR4455900 https://doi.org/10.1214/21-aoas1567

LIU, Y., DONG, Z., LIU, B., XU, Y. and DING, Z. (2023). FedForecast: A federated learning framework for short-term probabilistic individual load forecasting in smart grid. *Int. J. Electr. Power Energy Syst.* **152** 109172.

MA, J., ZHANG, J., LI, R., ZHENG, H. and LI, W. (2022). Using Bayesian optimization to automate the calibration of complex hydrological models: Framework and application. *Environ. Model. Softw.* **147** 105235.

MARTINEZ-VIOL, V., URBANO, E. M., DELGADO-PRIETO, M. and ROMERAL, L. (2022). Automatic model calibration for coupled HVAC and building dynamics using Modelica and Bayesian optimization. *Build. Environ.* **226** 109693.

MOČKUS, J. (1975). On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference* 400–404. Springer, Berlin.

MODELICA ASSOCIATION (2023). Modelica—a Unified Object-Oriented Language for Systems Modeling, Language Specification Version 3.6. https://specification.modelica.org/maint/3.6/MLS.pdf.

MUELLER FOUNDATION (2023). Mueller Affordable Homes Program: Frequently Asked Questions. https://muelleraustin.com/wp-content/uploads/2023/08/Mueller-Affordable-Homes-FAQ-1.pdf.

PALLONETTO, F., DE ROSA, M., D'ETTORRE, F. and FINN, D. P. (2020). On the assessment and control optimisation of demand response programs in residential buildings. *Renew. Sustain. Energy Rev.* **127** 109861.

PARK, J., BYON, E., KO, Y. M. and SHASHAANI, S. (2025). Strata design for variance reduction in stochastic simulation. *Technometrics* **67** 203–214. MR4902935 https://doi.org/10.1080/00401706.2024.2416411

PECAN STREET INC. (2025). Pecan Street Inc. https://www.pecanstreet.org/. Accessed: 2025-08-15.

RAMOS, P. V. B., VILLELA, S. M., SILVA, W. N. and DIAS, B. H. (2023). Residential energy consumption forecasting using deep learning models. *Appl. Energy* **350** 121705.

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

REGIS, R. G. (2016). Trust regions in Kriging-based optimization with expected improvement. *Eng. Optim.* **48** 1037–1059. MR3473811 https://doi.org/10.1080/0305215X.2015.1082350

REIKER, T., GOLUMBEANU, M., SHATTOCK, A., BURGERT, L., SMITH, T. A., FILIPPI, S., CAMERON, E. and PENNY, M. A. (2021). Emulator-based Bayesian optimization for efficient multi-objective calibration of an individual-based model of malaria. *Nat. Commun.* **12** 7212.

ROWAN, C. (2022). Pecan Street on KXAN: Why One Austin Neighborhood Tracks Energy Use down to the Circuit. https://www.pecanstreet.org/2022/07/kxan/.

RUSSO, D., VAN ROY, B., KAZEROUNI, A., OSBAND, I., WEN, Z. et al. (2018). A tutorial on Thompson sampling. *Found. Trends Mach. Learn.* **11** 1–96.

RUTTER, C. M., OZIK, J., DEYOREO, M. and COLLIER, N. (2019). Microsimulation model calibration using incremental mixture approximate Bayesian computation. *Ann. Appl. Stat.* **13** 2189–2212. MR4037427 https://doi.org/10.1214/19-aoas1279

SALEM, M. B., BACHOC, F., ROUSTANT, O., GAMBOA, F. and TOMASO, L. (2019). Gaussian process-based dimension reduction for goal-oriented sequential design. *SIAM/ASA J. Uncertain. Quantificat.* **7** 1369–1397. MR4041716 https://doi.org/10.1137/18M1167930

SALTELLI, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Comput. Phys. Commun.* **145** 280–297.

SALTELLI, A., TARANTOLA, S. and CHAN, K. P. S. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics* **41** 39–56.

SHA, D., OZBAY, K. and DING, Y. (2020). Applying Bayesian optimization for calibration of transportation simulation models. *Transp. Res. Rec.* **2674** 215–228.

SHEN, Y. and KINGSFORD, C. (2023). Computationally efficient high-dimensional Bayesian optimization via variable selection. In *Proceedings of the Second International Conference on Automated Machine Learning*. *Proceedings of Machine Learning Research* **224** 15/1–27.

SOBOL, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul.* **55** 271–280. MR1823119 https://doi.org/10.1016/S0378-4754(00)00270-6

SOKOL, J., DAVILA, C. C. and REINHART, C. F. (2017). Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. *Energy Build.* **134** 11–24.

SPAGNOL, A., LE RICHE, R. and DA VEIGA, S. (2019). Global sensitivity analysis for optimization with variable selection. *SIAM/ASA J. Uncertain. Quantificat.* **7** 417–443. MR3939340 https://doi.org/10.1137/18M1167978

SRINIVAS, N., KRAUSE, A., KAKADE, S. M. and SEEGER, M. W. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the* 27*th International Conference on Machine Learning* 1015–1022.

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

TUO, R. and WU, C. F. J. (2015). Efficient calibration for imperfect computer models. *Ann. Statist.* **43** 2331–2352. MR3405596 https://doi.org/10.1214/15-AOS1314

U. S. DEPARTMENT OF ENERGY (2019). EneryPlus essentials Technical Report.

U. S. ENERGY INFORMATION ADMINISTRATION (2015). Table 2.1 Energy Consumption by Sectors.

WINKEL, M. A., STALLRICH, J. W., STORLIE, C. B. and REICH, B. J. (2021). Sequential optimization in locally important dimensions. *Technometrics* **63** 236–248. MR4251497 https://doi.org/10.1080/00401706.2020.1714738

WYCOFF, N., BINOIS, M. and WILD, S. M. (2021). Sequential learning of active subspaces. *J. Comput. Graph. Statist.* **30** 1224–1237. MR4356616 https://doi.org/10.1080/10618600.2021.1874962

XU, Z., JEONG, C., BYON, E. and CETIN, K. (2021). Season-dependent parameter calibration in building energy simulation. In *Proceedings of* 2021 *IISE Annual Conference* 423–428.

YANG, Y., JI, C. and DENG, K. (2021). Rapid design of metamaterials via multitarget Bayesian optimization. *Ann. Appl. Stat.* **15** 768–796. MR4298963 https://doi.org/10.1214/20-aoas1426

YUAN, J. and NG, S. H. (2020). An integrated method for simultaneous calibration and parameter selection in computer models. *ACM Trans. Model. Comput. Simul.* **30** 1–23. MR4066848 https://doi.org/10.1145/3364217

ZHAN, S., WICHERN, G., LAUGHMAN, C., CHONG, A. and CHAKRABARTY, A. (2022). Calibrating building simulation models using multi-source datasets and meta-learned Bayesian optimization. *Energy Build.* **270** 112278.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1586–1604
https://doi.org/10.1214/25-AOAS2115

K-CONTACT DISTANCE FOR NOISY NONHOMOGENEOUS SPATIAL POINT DATA WITH APPLICATION TO REPEATING FAST RADIO BURST SOURCES

BY A. M. COOK[1,a], DAYI LI[2,d], GWENDOLYN M. EADIE[2,e], DAVID C. STENNING[4,i], PAUL SCHOLZ[5,k], DEREK BINGHAM[4,j], RADU CRAIU[2,f], B. M. GAENSLER[6,l], KIYOSHI W. MASUI[7,n], ZIGGY PLEUNIS[8,o], ANTONIO HERRERA-MARTIN[3,g], RONNIY C. JOSEPH[1,b], AYUSH PANDHI[3,h], AARON B. PEARLMAN[1,c] AND J. XAVIER PROCHASKA[6,m]

[1]*Department of Physics, McGill University,* [a]*amanda.cook@mail.mcgill.ca,* [b]*ronniy.joseph@mcgill.ca,* [c]*aaron.b.pearlman@physics.mcgill.ca*

[2]*Department of Statistical Science, University of Toronto,* [d]*dayi.li@mail.utoronto.ca,* [e]*gwen.eadie@utoronto.ca,* [f]*radu.craiu@utoronto.ca*

[3]*David A. Dunlap Department of Astronomy & Astrophysics, University of Toronto,* [g]*antonio.herreramartin@utoronto.ca,* [h]*ayush.pandhi@mail.utoronto.ca*

[4]*Department of Statistics and Actuarial Science, Simon Fraser University,* [i]*david_stenning@sfu.ca,* [j]*dbingham@sfu.ca*

[5]*Department of Physics and Astronomy, York University,* [k]*pscholz@yorku.ca*

[6]*Department of Astronomy and Astrophysics, University of California, Santa Cruz,* [l]*gaensler@ucsc.edu,* [m]*xavier@ucolick.org*

[7]*MIT Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology,* [n]*kmasui@mit.edu*

[8]*Anton Pannekoek Institute for Astronomy, University of Amsterdam,* [o]*z.pleunis@uva.nl*

This paper introduces an approach to analyze nonhomogeneous Poisson processes (NHPP) observed with noise, focusing on previously unstudied second-order characteristics of the noisy process. Utilizing a hierarchical Bayesian model with noisy data, we estimate hyperparameters governing a physically motivated NHPP intensity. Simulation studies demonstrate the reliability of this methodology in accurately estimating hyperparameters. Leveraging the posterior distribution, we then infer the probability of detecting a certain number of events within a given radius, the k-contact distance. We demonstrate our methodology with an application to observations of fast radio bursts (FRBs) detected by the Canadian Hydrogen Intensity Mapping Experiment's FRB Project (CHIME/FRB). This approach allows us to identify repeating FRB sources by bounding or directly simulating the probability of observing k physically independent sources within some radius in the detection domain or the *probability of coincidence* (P_C). The new methodology improves the repeater detection P_C in 91% of cases when applied to the largest sample of previously classified observations, with a median improvement factor (existing metric over P_C from our methodology) of ~ 4800.

REFERENCES

AGGARWAL, K., BUDAVÁRI, T., DELLER, A. T. et al. (2021). Probabilistic association of transients to their hosts (PATH). *Astrophys. J.* **911** 95.

ARVIZ DEVELOPERS (2020). ArviZ: Exploratory analysis of Bayesian models. Astrophysics Source Code Library, record ascl:2004.012.

BADDELEY, A., TURNER, R., MØLLER, J. and HAZELTON, M. (2005). Residual analysis for spatial point processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 617–666. MR2210685 https://doi.org/10.1111/j.1467-9868.2005.00519.x

BAPTISTA, J., PROCHASKA, J. X., MANNINGS, A. G. et al. (2024). Measuring the variance of the Macquart relation in redshift–extragalactic dispersion measure modeling. *Astrophys. J.* **965** 57.

Key words and phrases. Fast radio burst, noisy nonhomogeneous Poisson process, k-contact distance, spatial point processes, CHIME/FRB, hierarchical Bayesian modeling.

BAR-HEN, A., CHADŒUF, J., DESSARD, H. and MONESTIEZ, P. (2013). Estimating second order characteristics of point processes with known independent noise. *Stat. Comput.* **23** 297–309. MR3041437 https://doi.org/10.1007/s11222-011-9311-7

BECKWITH, S. V. W., STIAVELLI, M., KOEKEMOER, A. M. et al. (2006). The Hubble ultra deep field. *Astron. J.* **132** 1729–1755.

BERMAN, M. (1977). Distance distributions associated with Poisson processes of geometric figures. *J. Appl. Probab.* **14** 195–199. MR0438422 https://doi.org/10.2307/3213273

CHAKRABORTY, A. and GELFAND, A. E. (2010). Analyzing spatial point patterns subject to measurement error. *Bayesian Anal.* **5** 97–122. MR2596437 https://doi.org/10.1214/10-BA504

CHAWLA, P., KASPI, V. M., JOSEPHY, A. et al. (2017). A search for fast radio bursts with the GBNCC pulsar survey. *Astrophys. J.* **844** 140.

CHEN, B. H., HASHIMOTO, T., GOTO, T. et al. (2022). Uncloaking hidden repeating fast radio bursts with unsupervised machine learning. *Mon. Not. R. Astron. Soc.* **509** 1227–1236.

CHIME COLLABORATION et al. (2022). An overview of CHIME, the Canadian hydrogen intensity mapping experiment. *Astrophys. J., Suppl. Ser.* **261** 29.

CHIME/FRB COLLABORATION et al. (2018). The CHIME fast radio burst project: System overview. *Astrophys. J.* **863** 48.

CHIME/FRB COLLABORATION et al. (2021). The first CHIME/FRB fast radio burst catalog. *Astrophys. J., Suppl. Ser.* **257** 59.

CHIME/FRB COLLABORATION et al. (2023). CHIME/FRB discovery of 25 repeating fast radio burst sources. *Astrophys. J.* **947** 83.

COOK, A. M., BHARDWAJ, M., GAENSLER, B. M. et al. (2023). An FRB sent me a DM: Constraining the electron column of the Milky Way halo with fast radio burst dispersion measures from CHIME/FRB. *Astrophys. J.* **946** 58.

COOK, A. M., LI, D., EADIE, G. M. et al. (2026). Supplement to "K-contact distance for noisy nonhomogeneous spatial point data with application to repeating fast radio burst sources." https://doi.org/10.1214/25-AOAS2115SUPPA, https://doi.org/10.1214/25-AOAS2115SUPPB

CORDES, J. M. and LAZIO, T. J. W. (2002). NE2001.I. A new model for the Galactic distribution of free electrons and its fluctuations. arXiv e-prints. Available at arXiv:astro-ph/0207156.

CUCALA, L. (2008). Intensity estimation for spatial point processes observed with noise. *Scand. J. Stat.* **35** 322–334. MR2418744 https://doi.org/10.1111/j.1467-9469.2007.00583.x

DOLAG, K., GAENSLER, B. M., BECK, A. M. et al. (2015). Constraints on the distribution and energetics of fast radio bursts using cosmological hydrodynamic simulations. *Mon. Not. R. Astron. Soc.* **451** 4277–4289.

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

GLASSNER, A. S. (1990). *Graphics Gems. Graphics Gems—IBM*. Elsevier Science, Amsterdam.

HERRERA-MARTIN, A., CRAIU, R. V., EADIE, G. M. et al. (2025). Rare event classification with weighted logistic regression for identifying repeating fast radio bursts. *Astrophys. J.* **982** 46.

JAMES, C. W. (2023). Modelling repetition in zDM: A single population of repeating fast radio bursts can explain CHIME data. *Publ. Astron. Soc. Austral.* **40** e057.

JAMES, C. W., PROCHASKA, J. X., MACQUART, J. P. et al. (2022). The z-DM distribution of fast radio bursts. *Mon. Not. R. Astron. Soc.* **509** 4775–4802.

KIRSTEN, F., OULD-BOUKATTINE, O. S., HERRMANN, W. et al. (2024). A link between repeating and non-repeating fast radio bursts through their energy distributions. *Nat. Astron.* **8** 337–346.

KOTTAS, A. and SANSÓ, B. (2007). Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *J. Statist. Plann. Inference* **137** 3151–3163. MR2365118 https://doi.org/10.1016/j.jspi.2006.05.022

LANMAN, A. E., ANDREW, S., LAZDA, M. et al. (2024). CHIME/FRB Outriggers: KKO station system and commissioning results. arXiv e-prints. Available at arXiv:2402.07898.

LAWRENCE, E., VANDER WIEL, S., LAW, C. et al. (2017). The nonhomogeneous Poisson process for fast radio burst rates. *Astron. J.* **154** 117.

LI, D., WANG, P., ZHU, W. W. et al. (2021). A bimodal burst energy distribution of a repeating fast radio burst source. *Nature* **598** 267–271.

LUND, J., PENTTINEN, A. and RUDEMO, M. (1999). Bayesian analysis of spatial point patterns from noisy observations. Ph.D. thesis, Dept. Mathematics and Physics, The Royal Veterinary and Argicultural Univ., Copenhagen.

LUO, J.-W., ZHU-GE, J.-M. et al. (2023). Machine learning classification of CHIME fast radio bursts—I. Supervised methods. *Mon. Not. R. Astron. Soc.* **518** 1629–1641.

MACQUART, J. P., PROCHASKA, J. X., MCQUINN, M. et al. (2020). A census of baryons in the Universe from localized fast radio bursts. *Nature* **581** 391–395.

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

Moroianu, A., Wen, L., James, C. W. et al. (2023). An assessment of the association between a fast radio burst and binary neutron star merger. *Nat*. *Astron*. **7** 579–589.

Noutsos, A., Sobey, C., Kondratiev, V. I. et al. (2015). Pulsar polarisation below 200 MHz: Average profiles and propagation effects. *Astron*. *Astrophys*. **576** A62.

Petroff, E., Hessels, J. W. T. and Lorimer, D. R. (2022). Fast radio bursts at the dawn of the 2020s. *Astron*. *Astrophys*. *Rev.* **30** 2.

Platts, E., Weltman, A., Walters, A., Tendulkar, S. P., Gordin, J. E. B. and Kandhai, S. (2019). A living theory catalogue for fast radio bursts. *Phys*. *Rep*. **821** 1–27. MR4008577 https://doi.org/10.1016/j.physrep.2019.06.003

Pleunis, Z., Good, D. C., Kaspi, V. M. et al. (2021). Fast radio burst morphology in the first CHIME/FRB catalog. *Astrophys*. *J*. **923** 1.

Ravi, V., Catha, M., Chen, G. et al. (2023). Deep Synoptic Array science: a 50 Mpc fast radio burst constrains the mass of the Milky Way circumgalactic medium. arXiv e-prints. Available at arXiv:2301.01000.

Schechter, P. (1976). An analytic expression for the luminosity function for galaxies. *Astrophys*. *J*. **203** 297–306.

Shannon, R. M., Macquart, J. P., Bannister, K. W. et al. (2018). The dispersion-brightness relation for fast radio bursts from a wide-field survey. *Nature* **562** 386–390.

Spitler, L. G., Scholz, P., Hessels, J. W. T. et al. (2016). A repeating fast radio burst. *Nature* **531** 202–205.

van Lieshout, M. N. M. (2000). *Markov Point Processes and Their Applications*. Imperial College Press, London. MR1789230 https://doi.org/10.1142/9781860949760

Zhang, B. (2022). The physics of fast radio bursts. arXiv e-prints. Available at arXiv:2212.03972.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1605–1625
https://doi.org/10.1214/26-AOAS2158

ADAPTIVE BLOCK-BASED CHANGE-POINT DETECTION FOR SPARSE SPATIALLY CLUSTERED DATA WITH APPLICATIONS IN REMOTE SENSING IMAGING

BY ALAN MOORE[a], LYNNA CHU[b] AND ZHENGYUAN ZHU[c]

Department of Statistics, Iowa State University, [a]*alancm@iastate.edu,* [b]*lchu@iastate.edu,* [c]*zzhu@iastate.edu*

We present a nonparametric change-point detection approach to detect potentially sparse changes in a time series of high-dimensional observations or non-Euclidean data objects. We target a change in distribution that occurs in a small, unknown subset of dimensions, where these dimensions may be correlated. Our work is motivated by a remote sensing application, where changes occur in small, spatially clustered regions over time. An adaptive block-based change-point detection framework is proposed that accounts for spatial dependencies across dimensions and leverages these dependencies to boost detection power and improve estimation accuracy. Through simulation studies we demonstrate that our approach has superior performance in detecting sparse changes in datasets with spatial or local group structures. An application of the proposed method to detect activity, such as new construction, in remote sensing imagery of the Natanz Nuclear Facility in Iran is presented to demonstrate the method's efficacy.

REFERENCES

ALBRIGHT, D. and BURKHARD, S. (2022). Imagery Update: Iran Continues to Harden its New Natanz Tunnel Complex [1] | Institute for Science and International Security.

ALBRIGHT, D., BURKHARD, S. and FARAGASSO, S. (2024). Imagery Update: Construction is Ongoing at the Natanz Tunnel Facility | Institute for Science and International Security.

BARDWELL, L., FEARNHEAD, P., ECKLEY, I. A., SMITH, S. and SPOTT, M. (2019). Most recent changepoint detection in panel data. *Technometrics* **61** 88–98. https://doi.org/10.1080/00401706.2018.1438926

BHATTACHARJEE, M., BANERJEE, M. and MICHAILIDIS, G. (2019). Change Point Estimation in Panel Data with Temporal and Cross-sectional Dependence. Publisher: arXiv Version Number: 1. https://doi.org/10.48550/ARXIV.1904.11101

BHATTACHARYA, B. B. (2020). Asymptotic distribution and detection thresholds for two-sample tests based on geometric graphs. *Ann. Statist.* **48** 2879–2903. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/19-AOS1913

CAI, H. and WANG, T. (2023). Estimation of high-dimensional change-points under a group sparsity structure. *Electron. J. Stat.* **17** 858–894. Publisher: Institute of Mathematical Statistics and Bernoulli Society. https://doi.org/10.1214/23-EJS2116

CHEN, H. and CHU, L. (2023). Graph-based change-point analysis. *Annu. Rev. Stat. Appl.* **10** 475–499. https://doi.org/10.1146/annurev-statistics-122121-033817

CHO, H. (2016). Change-point detection in panel data via double CUSUM statistic. *Electron. J. Stat.* **10** 2000–2038. Publisher: Institute of Mathematical Statistics and Bernoulli Society. https://doi.org/10.1214/16-EJS1155

CHU, L. and CHEN, H. (2019). Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data. *Ann. Statist.* **47**. https://doi.org/10.1214/18-AOS1691

DUBEY, P. and MÜLLER, H.-G. (2020). Fréchet change-point detection. *Ann. Statist.* **48**. https://doi.org/10.1214/19-AOS1930

ENIKEEVA, F. and HARCHAOUI, Z. (2019). High-dimensional change-point detection under sparse alternatives. *Ann. Statist.* **47**. https://doi.org/10.1214/18-AOS1740

FRIEDMAN, J. H. and RAFSKY, L. C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7**. https://doi.org/10.1214/aos/1176344722

Key words and phrases. Change-point, nonparametric, graph-based tests, spatial dependence, high-dimensional data, satellite images.

FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42** 2243–2281. Publisher: Institute of Mathematical Statistics.

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

HOLLAWAY, M. J. and KILLICK, R. (2024). Detection of spatiotemporal changepoints: A generalised additive model approach. *Stat. Comput.* **34** 162. https://doi.org/10.1007/s11222-024-10478-6

JIANG, F., ZHU, C. and SHAO, X. (2024). Two-sample and change-point inference for non-Euclidean valued time series. *Electron. J. Stat.* **18** 848–894. Publisher: Institute of Mathematical Statistics and Bernoulli Society. https://doi.org/10.1214/24-EJS2218

JIRAK, M. (2015). Uniform change point tests in high dimension. *Ann. Statist.* **43** 2451–2483. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/15-AOS1347

KOVÁCS, S., BÜHLMANN, P., LI, H. and MUNK, A. (2023). Seeded binary segmentation: A general methodology for fast and optimal changepoint detection. *Biometrika* **110** 249–256. https://doi.org/10.1093/biomet/asac052

KULLDORFF, M. (1997). A spatial scan statistic. *Comm. Statist. Theory Methods* **26** 1481–1496. https://doi.org/10.1080/03610929708831995

KULLDORFF, M., HUANG, L. and KONTY, K. (2009). A scan statistic for continuous data based on the normal probability model. *Int. J. Health Geogr.* **8** 58. https://doi.org/10.1186/1476-072X-8-58

LI, J., CHEN, L., WANG, W. and WU, W. B. (2024). l2 inference for change points in high-dimensional time series via a two-way MOSUM. *Ann. Statist.* **52** 602–627. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/24-AOS2360

LIU, B., ZHOU, C., ZHANG, X. and LIU, Y. (2020). A unified data-adaptive framework for high dimensional change point detection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 933–963. https://doi.org/10.1111/rssb.12375

LIU, Y.-W. and CHEN, H. (2022). A fast and efficient change-point detection framework based on approximate k-nearest neighbor graphs. *IEEE Trans. Signal Process.* **70** 1976–1986. Conference Name: IEEE Transactions on Signal Processing. https://doi.org/10.1109/TSP.2022.3162120

LOH, J. M. and ZHU, Z. (2007). Accounting for spatial correlation in the scan statistic. *Ann. Appl. Stat.* **1**. https://doi.org/10.1214/07-AOAS129

MATTESON, D. S. and JAMES, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *J. Amer. Statist. Assoc.* **109** 334–345. Publisher: Taylor & Francis. https://doi.org/10.1080/01621459.2013.849605

MOORE, A., CHU, L. and ZHU, Z. (2026). Supplement to "Adaptive block-based change-point detection for sparse spatially clustered data with applications in remote sensing imaging." https://doi.org/10.1214/26-AOAS2158SUPPA, https://doi.org/10.1214/26-AOAS2158SUPPB

MORADI, M., CRONIE, O., PÉREZ-GOYA, U. and MATEU, J. (2023). Hierarchical spatio-temporal change-point detection. *Amer. Statist.* **77** 390–400. https://doi.org/10.1080/00031305.2023.2191670

SCOTT, A. J. and KNOTT, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* **30** 507. https://doi.org/10.2307/2529204

WANG, T. and SAMWORTH, R. J. (2018). High dimensional change point estimation via sparse projection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 57–83. https://doi.org/10.1111/rssb.12243

ZHANG, L. and ZHU, Z. (2012). Spatial Multiresolution Cluster Detection Method. Available at arXiv:1205.2106 [stat]. https://doi.org/10.48550/arXiv

ZHANG, Y., WANG, R. and SHAO, X. (2022). Adaptive inference for change points in high-dimensional data. *J. Amer. Statist. Assoc.* **117** 1751–1762. https://doi.org/10.1080/01621459.2021.1884562

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1626–1645
https://doi.org/10.1214/26-AOAS2199

HIERARCHICAL PROBABILISTIC CONFORMAL PREDICTION FOR DISTRIBUTED ENERGY RESOURCES ADOPTION

BY WENBIN ZHOU[a] AND SHIXIANG ZHU[b]

Heinz College of Information Systems and Public Policy, Carnegie Mellon University, [a]*wenbinz2@andrew.cmu.edu,* [b]*shixiangzhu@cmu.edu*

The rapid growth of distributed energy resources (DERs) presents both opportunities and operational challenges for electric grid management. Accurately predicting DER adoption is critical for proactive infrastructure planning, but the inherent uncertainty and spatial disparity of DER growth complicate traditional forecasting approaches. Moreover, the hierarchical structure of distribution grids demands that predictions satisfy statistical guarantees at both the circuit and substation levels, a nontrivial requirement for reliable decision-making. In this paper we propose a novel uncertainty quantification framework for DER adoption predictions that ensures validity across hierarchical grid structures. Leveraging a multivariate Hawkes process to model DER adoption dynamics and a tailored split conformal prediction algorithm, we introduce a new nonconformity score that preserves statistical guarantees under aggregation while maintaining prediction efficiency. We establish theoretical validity under mild conditions and, through empirical evaluation on customer-level solar panel installation data from Indianapolis, Indiana, demonstrate that our method consistently outperforms existing baselines in both predictive accuracy and uncertainty calibration.

REFERENCES

ALMEIDA, V., RIBEIRO, R. and GAMA, J. (2016). Hierarchical time series forecast in electrical grids. In *Information Science and Applications* (*ICISA*) **2016** 995–1005. Springer, Berlin.

ANGELOPOULOS, A. N. and BATES, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint. Available at arXiv:2107.07511.

BAGHERI, A., ZHAO, C., QIU, F. and WANG, J. (2018). Resilient transmission hardening planning in a high renewable penetration era. *IEEE Trans. Power Syst.* **34** 873–882.

BAHERI, A. and SHAHBAZI, M. A. (2025). Multi-scale conformal prediction: A theoretical framework with coverage guarantees. arXiv preprint. Available at arXiv:2502.05565.

BARBER, R. F., CANDÈS, E. J., RAMDAS, A. and TIBSHIRANI, R. J. (2023). Conformal prediction beyond exchangeability. *Ann. Statist.* **51** 816–845. MR4601003 https://doi.org/10.1214/23-aos2276

BARBOSE, G. L., FORRESTER, S., O'SHAUGHNESSY, E. and DARGHOUTH, N. R. (2022). Residential solar-adopter income and demographic trends: 2022 update.

BERNARDS, R., MORREN, J. and SLOOTWEG, H. (2018). Development and implementation of statistical models for estimating diversified adoption of energy transition technologies. *IEEE Trans. Sustain. Energy* **9** 1540–1554.

BOLLINGER, B., DARGHOUTH, N., GILLINGHAM, K. and GONZALEZ-LIRA, A. (2024a). Valuing technology complementarities: Rooftop solar and energy storage Technical Report, National Bureau of Economic Research.

BOLLINGER, B. and GILLINGHAM, K. (2012). Peer effects in the diffusion of solar photovoltaic panels. *Mark. Sci.* **31** 900–912.

BOLLINGER, B., GILLINGHAM, K., LAMP, S. and TSVETANOV, T. (2024b). Promotional campaign duration and word of mouth in solar panel adoption. *Mark. Sci.* **43** 1132–1148.

U. S. CENSUS BUREAU (2025). data.census.gov. https://data.census.gov. Accessed 2025-04-15.

DHARSHING, S. (2017). Household dynamics of technology adoption: A spatial econometric analysis of residential solar photovoltaic (PV) systems in Germany. *Energy Res. Soc. Sci.* **23** 113–124.

Key words and phrases. Conformal prediction, multivariate temporal point processes, hierarchical prediction, distributed energy resources.

DHEUR, V., BOSSER, T., IZBICKI, R. and BEN TAIEB, S. (2024). Distribution-free conformal joint prediction regions for neural marked temporal point processes. *Mach. Learn.* **113** 7055–7102. MR4783186 https://doi.org/10.1007/s10994-024-06594-z

DONG, Z., ZHU, S., XIE, Y., MATEU, J. and RODRÍGUEZ-CORTÉS, F. J. (2023). Non-stationary spatio-temporal point process modeling for high-resolution COVID-19 data. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **72** 368–386. MR4724080 https://doi.org/10.1093/jrsssc/qlad013

FOTHERINGHAM, A. S. and WONG, D. W. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environ. Plan. A* **23** 1025–1044.

GEORGILAKIS, P. S. and HATZIARGYRIOU, N. D. (2015). A review of power distribution planning in the modern power systems era: Models, methods and future research. *Electr. Power Syst. Res.* **121** 89–100.

GIBBS, I. and CANDES, E. (2021). Adaptive conformal inference under distribution shift. *Adv. Neural Inf. Process. Syst.* **34** 1660–1672.

GILLINGHAM, K. T. and BOLLINGER, B. (2021). Social learning and solar photovoltaic adoption. *Manag. Sci.* **67** 7091–7112.

GUHA, S. and KUMAR, S. (2018). Emergence of big data research in operations management, information systems, and healthcare: Past contributions and future roadmap. *Prod. Oper. Manag.* **27** 1724–1735.

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

HOROWITZ, K. A., PETERSON, Z., CODDINGTON, M. H., DING, F., SIGRIN, B. O., SALEEM, D., BALDWIN, S. E., LYDIC, B., STANFIELD, S. C. et al. (2019). An overview of distributed energy resource (DER) interconnection: Current practices and emerging solutions.

HYNDMAN, R. J., AHMED, R. A., ATHANASOPOULOS, G. and SHANG, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Comput. Statist. Data Anal.* **55** 2579–2589. MR2802337 https://doi.org/10.1016/j.csda.2011.03.006

IZBICKI, R., SHIMIZU, G. and STERN, R. B. (2022). `CD-split` and `HPD-split`: Efficient conformal regions in high dimensions. *J. Mach. Learn. Res.* **23** 87. MR4576672

JUNG, A.-H., LEE, D.-H., KIM, J.-Y., KIM, C. K., KIM, H.-G. and LEE, Y.-S. (2022). Regional photovoltaic power forecasting using vector autoregression model in South Korea. *Energies* **15** 7853.

KHATOR, S. K. and LEUNG, L. C. (1997). Power distribution planning: A review of models and issues. *IEEE Trans. Power Syst.* **12** 1151–1159.

KOUVELIS, P., CHAMBERS, C. and WANG, H. (2006). Supply chain management research and production and operations management: Review, trends, and opportunities. *Prod. Oper. Manag.* **15** 449–469.

LAURET, P., DAVID, M. and PEDRO, H. (2017). Probabilistic solar forecasting using quantile regression models. *Energies* **10**.

LEI, J., ROBINS, J. and WASSERMAN, L. (2013). Distribution-free prediction sets. *J. Amer. Statist. Assoc.* **108** 278–287. MR3174619 https://doi.org/10.1080/01621459.2012.751873

LEI, J. and WASSERMAN, L. (2014). Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 71–96. MR3153934 https://doi.org/10.1111/rssb.12021

LI, Y., SUN, Y., WANG, Q., SUN, K., LI, K.-J. and ZHANG, Y. (2023). Probabilistic harmonic forecasting of the distribution system considering time-varying uncertainties of the distributed energy resources and electrical loads. *Appl. Energy* **329** 120298.

MACK, R., SAKIB, M. and SUCCAR, S. (2017). Impacts of substation transformer backfeed at high PV penetrations. In 2017 *IEEE Power & Energy Society General Meeting* 1–5. IEEE.

MOHAMMED, N. A. and AL-BAZI, A. (2021). Management of renewable energy production and distribution planning using agent-based modelling. *Renew. Energy* **164** 509–520.

NOVOA, L., FLORES, R. and BROUWER, J. (2019). Optimal renewable generation and battery storage sizing and siting considering local transformer limits. *Appl. Energy* **256** 113926.

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

PAPADOPOULOS, H., PROEDROU, K., VOVK, V. and GAMMERMAN, A. (2002). Inductive confidence machines for regression. In *Machine Learning*: *ECML* 2002: 13*th European Conference on Machine Learning*. *Proceedings* **13** 345–356. Springer, Berlin.

PRINCIPATO, G., STOLTZ, G., AMARA-OUALI, Y., GOUDE, Y., HAMROUCHE, B. and POGGI, J.-M. (2024). Conformal prediction for hierarchical data. arXiv preprint, arXiv:2411.13479.

QUAN, H., KHOSRAVI, A., YANG, D. and SRINIVASAN, D. (2019). A survey of computational intelligence techniques for wind power uncertainty quantification in smart grids. *IEEE Trans. Neural Netw. Learn. Syst.* **31** 4582–4599.

REINHART, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statist. Sci.* **33** 299–318. MR3843374 https://doi.org/10.1214/17-STS629

SÁEZ, D., ÁVILA, F., OLIVARES, D., CAÑIZARES, C. and MARÍN, L. (2014). Fuzzy prediction interval models for forecasting renewable resources and loads in microgrids. *IEEE Trans. Smart Grid* **6** 548–556.

SHAFER, G. and VOVK, V. (2008). A tutorial on conformal prediction. *J. Mach. Learn. Res.* **9** 371–421. MR2417240

SHI, Z., LIANG, H. and DINAVAHI, V. (2017). Direct interval forecast of uncertain wind power based on recurrent neural networks. *IEEE Trans. Sustain. Energy* **9** 1177–1187.

SOLAR ENERGY INDUSTRIES ASSOCIATION (SEIA) (2023). Solar Market Insight Report 2023, Accessed: April 2025..

STANFIELD, S., ZAKAI, Y. and MCKERLEY, M. (2021). Key decisions for hosting capacity analyses. In *IREC* **37**.

SUN, S. H. and YU, R. (2023). Copula conformal prediction for multi-step time series prediction. In *The Twelfth International Conference on Learning Representations*.

TIBSHIRANI, R. J., FOYGEL BARBER, R., CANDES, E. and RAMDAS, A. (2019). Conformal prediction under covariate shift. *Adv. Neural Inf. Process. Syst.* **32**.

TUMU, R., CLEAVELAND, M., MANGHARAM, R., PAPPAS, G. J. and LINDEMANN, L. (2023). Multi-modal conformal prediction regions by optimizing convex shape templates. arXiv preprint, arXiv:2312.07434.

VAN DER MEER, D. W., SHEPERO, M., SVENSSON, A., WIDÉN, J. and MUNKHAMMAR, J. (2018). Probabilistic forecasting of electricity consumption, photovoltaic power generation and net demand of an individual building using Gaussian processes. *Appl. Energy* **213** 195–207.

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York. MR2161220

WANG, H. and SUN, B. (2021). Diffusion mechanism of leading technology in the new energy industry based on the bass model. *Front. Energy Res.* **9** 586787.

WANG, K., QI, X. and LIU, H. (2019). Photovoltaic power forecasting based LSTM-convolutional network. *Energy* **189** 116225.

WANG, Z., GAO, R., YIN, M., ZHOU, M. and BLEI, D. (2023). Probabilistic conformal prediction using conditional random samples. In *International Conference on Artificial Intelligence and Statistics* 8814–8836. PMLR.

WASSERMAN, S. and FAUST, K. (1994). *Social Network Analysis*: *Methods and Applications*.

WILLEMS, N., SEKAR, A., SIGRIN, B. and RAI, V. (2022). Forecasting distributed energy resources adoption for power systems. *iScience* **25** 104381.

WILLIAMS, E., CARVALHO, R., HITTINGER, E. and RONNENBERG, M. (2020). Empirical development of parsimonious model for international diffusion of residential solar. *Renew. Energy* **150** 570–577.

WOOD MACKENZIE (2023). Energy Transition Outlook 2023: Highlights. Accessed: 2025-04-28.

WU, X., CONEJO, A. J. and MATHEW, S. (2020). Optimal siting of batteries in distribution systems to enhance reliability. *IEEE Trans. Power Deliv.* **36** 3118–3127.

XU, C., JIANG, H. and XIE, Y. (2024). Conformal prediction for multi-dimensional time series by ellipsoidal sets. In *Forty-First International Conference on Machine Learning*.

XU, C., SUN, Y., DU, A. and GAO, D.-C. (2023). Quantile regression based probabilistic forecasting of renewable energy generation and building electrical load: A state of the art review. *J. Build. Eng.* **107772**.

XU, C. and XIE, Y. (2021). Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning* 11559–11569. PMLR.

XU, C. and XIE, Y. (2023). Sequential predictive conformal inference for time series. In *International Conference on Machine Learning* 38707–38727. PMLR.

ZHANG, H., VOROBEYCHIK, Y., LETCHFORD, J. and LAKKARAJU, K. (2016). Data-driven agent-based modeling, with application to rooftop solar adoption. *Auton. Agents Multi-Agent Syst.* **30** 1023–1049.

ZHANG, Y., WANG, J. and LI, Z. (2019). Uncertainty modeling of distributed energy resources: Techniques and challenges. *Curr. Sustain./Renew. Energy Rep.* **6** 42–51.

ZHENG, M. and ZHU, S. (2024). Generative conformal prediction with vectorized non-conformity scores. arXiv preprint. Available at arXiv:2410.13735.

ZHU, S., DING, R., ZHANG, M., VAN HENTENRYCK, P. and XIE, Y. (2021a). Spatio-temporal point processes with attention for traffic congestion event modeling. *IEEE Trans. Intell. Transp. Syst.* **23** 7298–7309.

ZHOU, W. and ZHU, S. (2026). Supplement to "Hierarchical probabilistic conformal prediction for distributed energy resources adoption." https://doi.org/10.1214/26-AOAS2199SUPPA, https://doi.org/10.1214/26-AOAS2199SUPPB

ZHU, S. and XIE, Y. (2022). Spatiotemporal-textual point processes for crime linkage detection. *Ann. Appl. Stat.* **16** 1151–1170. MR4438828 https://doi.org/10.1214/21-aoas1538

ZHU, S., YAO, R., XIE, Y., QIU, F., QIU, Y. and WU, X. (2021b). Quantifying grid resilience against extreme weather using large-scale customer power outage data. arXiv preprint. Available at arXiv:2109.09711.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1646–1670
https://doi.org/10.1214/25-AOAS2120

SPATIOTEMPORAL-NETWORK POINT PROCESSES FOR MODELING CRIME EVENTS WITH LANDMARKS

BY ZHENG DONG[1,a], JORGE MATEU[2,c] AND YAO XIE[1,b]

[1]*H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology,* [a]*zdong76@gatech.edu*, [b]*yao.xie@isye.gatech.edu*

[2]*Department of Mathematics, Universitat Jaume I,* [c]*mateu@mat.uji.es*

Self-exciting point processes are widely used to model the contagious effects of crime events living within continuous geographic space, using their occurrence time and locations. However, in urban environments most events are naturally constrained within the city's street network structure, and the contagious effects of crime are governed by such a network geography. Meanwhile, the complex distribution of urban infrastructures also plays an important role in shaping crime patterns across space. We introduce a novel spatiotemporal-network point process framework for crime modeling that integrates these urban environmental characteristics by incorporating self-attention graph neural networks. Our framework incorporates the street network structure as the underlying event space, where crime events can occur at random locations on the network edges. To realistically capture criminal movement patterns, distances between events are measured using street network distances. We then propose a new mark for a crime event by concatenating the event's crime category with the type of its nearby landmark, aiming to capture how the urban design influences the mixing structures of various crime types. A graph attention network architecture is adopted to learn the existence of mark-to-mark interactions. Extensive experiments on crime data from Valencia, Spain, demonstrate the effectiveness of our framework in understanding the crime landscape and forecasting crime risks across regions.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19** 716–723.

AKAIKE, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike* 199–213. Springer, Berlin.

ANDRESEN, M. A. (2007). Location quotients, ambient populations, and the spatial analysis of crime in Vancouver, Canada. *Environ. Plan. A* **39** 2423–2444.

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

BONTA, J., LAW, M. and HANSON, K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: A meta-analysis. *Psychol. Bull.* **123** 123.

BOUNCE (2024). Is Valencia safe to visit? A comprehensive safety guide.

BOWERS, K. J., JOHNSON, S. D. and PEASE, K. (2004). Prospective hot-spotting: The future of crime mapping? *Br. J. Criminol.* **44** 641–658.

BROWNING, C. R., BYRON, R. A., CALDER, C. A., KRIVO, L. J., KWAN, M.-P., LEE, J.-Y. and PETERSON, R. D. (2010). Commercial density, residential concentration, and crime: Land use patterns and violence in neighborhood context. *J. Res. Crime Delinq.* **47** 329–357.

CAI, B., ZHANG, J. and GUAN, Y. (2024). Latent network structure learning from high-dimensional multivariate point processes. *J. Amer. Statist. Assoc.* **119** 95–108. MR4713878 https://doi.org/10.1080/01621459.2022.2102019

CHAINEY, S., TOMPSON, L. and UHLIG, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Secur. J.* **21** 4–28.

CHENG, X., DONG, Z. and XIE, Y. (2025). Deep spatio-temporal point processes: Advances and new directions. arXiv preprint. Available at arXiv:2504.06364.

Key words and phrases. Spatiotemporal-network point process, crime data modeling, street networks with landmarks.

CHO, Y.-S., GALSTYAN, A., BRANTINGHAM, P. J. and TITA, G. (2014). Latent self-exciting point process model for spatial-temporal networks. *Discrete Contin. Dyn. Syst. Ser. B* **19** 1335–1354. MR3199782 https://doi.org/10.3934/dcdsb.2014.19.1335

OPENSTREETMAP CONTRIBUTORS (2017). Planet dump retrieved from https://planet.osm.org. https://www.openstreetmap.org.

D'ANGELO, N., PAYARES, D., ADELFIO, G. and MATEU, J. (2024). Self-exciting point process modelling of crimes on linear networks. *Stat. Model.* **24** 139–168. MR4715729 https://doi.org/10.1177/1471082X221094146

DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes. Vol. I: Elementary Theory and Methods*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR1950431

DONG, Z., CHENG, X. and XIE, Y. (2023). Spatio-temporal point processes with deep non-stationary kernels. In *The Eleventh International Conference on Learning Representations*.

DONG, Z., MATEU, J. and XIE, Y. (2026). Supplement to "Spatiotemporal-network point processes for modeling crime events with landmarks." https://doi.org/10.1214/25-AOAS2120SUPPA, https://doi.org/10.1214/25-AOAS2120SUPPB

DONG, Z., REPASKY, M., CHENG, X. and XIE, Y. (2023a). Deep graph kernel point processes. In *Temporal Graph Learning Workshop @ NeurIPS* 2023.

DONG, Z. and XIE, Y. (2024). Atlanta gun violence modeling via nonstationary spatio-temporal point processes. arXiv preprint. Available at arXiv:2408.09258.

DONG, Z., ZHU, S., XIE, Y., MATEU, J. and RODRÍGUEZ-CORTÉS, F. J. (2023b). Non-stationary spatio-temporal point process modeling for high-resolution COVID-19 data. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **72** 368–386. MR4724080 https://doi.org/10.1093/jrsssc/qlad013

DU, N., DAI, H., TRIVEDI, R., UPADHYAY, U., GOMEZ-RODRIGUEZ, M. and SONG, L. (2016). Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1555–1564.

FANG, G., XU, G., XU, H., ZHU, X. and GUAN, Y. (2024). Group network Hawkes process. *J. Amer. Statist. Assoc.* **119** 2328–2344. MR4797944 https://doi.org/10.1080/01621459.2023.2257889

FLEMING, Z., BRANTINGHAM, P., BRANTINGHAM, P. et al. (1994). Exploring auto theft in British Columbia. *Crime Prev. Stud.* **3** 47–90.

GAO, S., JANOWICZ, K. and COUCLELIS, H. (2017). Extracting urban functional regions from points of interest and human activities on location-based social networks. *Trans. GIS* **21** 446–467.

GOTTFREDSON, M. R. (1981). On the etiology of criminal victimization. *J. Crim. Law Criminol.* **72** 714.

GRANN, M., LÅNGSTRÖM, N., TENGSTRÖM, A. and KULLGREN, G. (1999). Psychopathy (PCL-R) predicts violent recidivism among criminal offenders with personality disorders in Sweden. *Law Hum. Behav.* **23** 205–217.

GROFF, E. (2011). Exploring 'near': Characterizing the spatial extent of drinking place influence on crime. *Aust. N. Z. J. Criminol.* **44** 156–179.

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

HESSELLUND, K. B., XU, G., GUAN, Y. and WAAGEPETERSEN, R. (2022a). Second-order semi-parametric inference for multivariate log Gaussian Cox processes. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **71** 244–268. MR4376853 https://doi.org/10.1111/rssc.12530

HESSELLUND, K. B., XU, G., GUAN, Y. and WAAGEPETERSEN, R. (2022b). Semiparametric multinomial logistic regression for multivariate point pattern data. *J. Amer. Statist. Assoc.* **117** 1500–1515. MR4480727 https://doi.org/10.1080/01621459.2020.1863812

HU, Y. and HAN, Y. (2019). Identification of urban functional areas based on poi data: A case study of the guangzhou economic and technological development zone. *Sustainability* **11** 1385.

JIANG, S., ALVES, A., RODRIGUES, F., FERREIRA, J. JR and PEREIRA, F. C. (2015). Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Comput. Environ. Urban Syst.* **53** 36–46.

JOHNSON, S. D. (2008). Repeat burglary victimisation: A tale of two theories. *Journal of Experimental Criminology* **4** 215–240.

JOHNSON, S. D. and BOWERS, K. J. (2010). Permeability and burglary risk: Are cul-de-sacs safer? *J. Quant. Criminol.* **26** 89–111.

KENNEDY, L. W., CAPLAN, J. M. and PIZA, E. (2011). Risk clusters, hotspots, and spatial intelligence: Risk terrain modeling as an algorithm for police resource allocation strategies. *J. Quant. Criminol.* **27** 339–362.

KENNEDY, L. W., CAPLAN, J. M., PIZA, E. L. and BUCCINE-SCHRAEDER, H. (2016). Vulnerability and exposure to crime: Applying risk terrain modeling to the study of assault in Chicago. *Appl. Spat. Anal. Policy* **9** 529–548.

KINNEY, J. B., BRANTINGHAM, P. L., WUSCHKE, K., KIRK, M. G. and BRANTINGHAM, P. J. (2008). Crime attractors, generators and detractors: Land use and urban crime opportunities. *Built Environ.* **34** 62–74.

KUMAZAWA, T. and OGATA, Y. (2014). Nonstationary ETAS models for nonstandard earthquakes. *Ann. Appl. Stat.* **8** 1825–1852. MR3271355 https://doi.org/10.1214/14-AOAS759

LEV-WIESEL, R., AMIR, M. and BESSER, A. (2004). Posttraumatic growth among female survivors of childhood sexual abuse in relation to the perpetrator identity. *J. Loss Trauma* **10** 7–17.

LEVINE, N. and CRIMESTAT, I. (2002). A spatial statistics program for the analysis of crime incident locations. Ned Levine and Associates, Houston, TX, and the National Institute of Justice, Washington, DC.

LEWIS, E., MOHLER, G., BRANTINGHAM, P. J. and BERTOZZI, A. L. (2012). Self-exciting point process models of civilian deaths in Iraq. *Secur. J.* **25** 244–264.

LINDERMAN, S. and ADAMS, R. (2014). Discovering latent network structure in point process data. In *International Conference on Machine Learning* 1413–1421. PMLR.

LIU, X., CARTER, J., RAY, B. and MOHLER, G. (2021). Point process modeling of drug overdoses with heterogeneous and missing data. *Ann. Appl. Stat.* **15** 88–101. MR4255268 https://doi.org/10.1214/20-aoas1384

LOEFFLER, C. and FLAXMAN, S. (2018). Is gun violence contagious? A spatiotemporal test. *J. Quant. Criminol.* **34** 999–1017.

LONG, Y., SHEN, Z., LONG, Y. and SHEN, Z. (2015). Discovering functional zones using bus smart card data and points of interest in Beijing. *Geosp. Anal. Support Urban Plan. Beijing* 193–217.

MAAS, A. L., HANNUN, A. Y., NG, A. Y. et al. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML* **30** 3, Atlanta, GA.

MEERA, A. K. and JAYAKUMAR, M. D. (1995). Determinants of crime in a developing country: A regression model. *Appl. Econ.* **27** 455–460.

MEI, H. and EISNER, J. M. (2017). The neural Hawkes process: A neurally self-modulating multivariate point process. *Adv. Neural Inf. Process. Syst.* **30**.

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

MOHLER, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast.* **30** 491–497.

MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. MR2816705 https://doi.org/10.1198/jasa.2011.ap09546

MØLLER, J. and WAAGEPETERSEN, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. *Monographs on Statistics and Applied Probability* **100**. CRC Press/CRC, Boca Raton, FL. MR2004226

NEILL, D. B. and GORR, W. L. (2007). Detecting and preventing emerging epidemics of crime. *Adv. Dis. Surveill.* **4**.

NEWMAN, M. E. J. (2010). *Networks*: *An Introduction*. Oxford Univ. Press, Oxford. MR2676073 https://doi.org/10.1093/acprof:oso/9780199206650.001.0001

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

OGATA, Y. (1998). Space-time point-process models for earthquake occurrences. *Ann. Inst. Statist. Math.* **50** 379–402.

PERRY, W. L. (2013). Predictive policing: The role of crime forecasting in law enforcement operations. *Rand Corp.*

PORTER, M. D. and WHITE, G. (2012). Self-exciting hurdle models for terrorist activity. *Ann. Appl. Stat.* **6** 106–124. MR2951531 https://doi.org/10.1214/11-AOAS513

REINHART, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statist. Sci.* **33** 299–318. MR3843374 https://doi.org/10.1214/17-STS629

REINHART, A. and GREENHOUSE, J. (2018). Self-exciting point processes with spatial covariates: Modelling the dynamics of crime. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **67** 1305–1329. MR3873709 https://doi.org/10.1111/rssc.12277

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

ROSSMO, D. K. (1999). *Geographic Profiling*. CRC Press, Boca Raton.

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

RUSSO, S., ROCCATO, M. and VIENO, A. (2013). Criminal victimization and crime risk perception: A multilevel longitudinal study. *Soc. Indic. Res.* **112** 535–548.

SANNA PASSINO, F., CHE, Y. and CARDOSO CORREIA PERELLO, C. (2024). Graph-based mutually exciting point processes for modelling event times in docked bike-sharing systems. *Stat* **13** e660. MR4719011 https://doi.org/10.1002/sta4.660

SHCHUR, O., TÜRKMEN, A. C., JANUSCHOWSKI, T. and GÜNNEMANN, S. (2021). Neural temporal point processes: a review. arXiv preprint. Available at arXiv:2104.03528.

SHORT, M. B., D'ORSOGNA, M. R., PASOUR, V. B., TITA, G. E., BRANTINGHAM, P. J., BERTOZZI, A. L. and CHAYES, L. B. (2008). A statistical model of criminal behavior. *Math. Models Methods Appl. Sci.* **18** 1249–1267. MR2438215 https://doi.org/10.1142/S0218202508003029

STUCKY, T. D. and OTTENSMANN, J. R. (2009). Land use and violent crime. *Criminology* **47** 1223–1264.

TARZIA, L., THURAISINGAM, S., NOVY, K., VALPIED, J., QUAKE, R. and HEGARTY, K. (2018). Exploring the relationships between sexual violence, mental health and perpetrator identity: A cross-sectional Australian primary care study. *BMC Public Health* **18** 1–9.

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

VELIČKOVIĆ, P., CUCURULL, G., CASANOVA, A., ROMERO, A., LIÒ, P. and BENGIO, Y. (2018). Graph attention networks. In *International Conference on Learning Representations*.

WANG, X. and BROWN, D. E. (2012). The spatio-temporal modeling for criminal incidents. *Secur. Inform.* **1** 1–17.

WEI, N., WALTEROS, J. L. and BATTA, R. (2020). On the distance between random events on a network. *Networks* **75** 203–231. MR4066840 https://doi.org/10.1002/net.21919

WEISBURD, D., GROFF, E. R. and YANG, S.-M. (2012). *The Criminology of Place*: *Street Segments and Our Understanding of the Crime Problem*. Oxford Univ. Press, London.

WU, W., LIU, H., ZHANG, X., LIU, Y. and ZHA, H. (2020). Modeling event propagation via graph biased temporal point process. *IEEE Trans. Neural Netw. Learn. Syst.*

WUSCHKE, K. and KINNEY, J. B. (2018). 475Built environment, land use, and crime. In *The Oxford Handbook of Environmental Criminology* Oxford Univ. Press, London.

XIA, W., LI, Y. and LI, S. (2022). Graph neural point process for temporal interaction prediction. *IEEE Trans. Knowl. Data Eng.*

XU, G., LIANG, C., WAAGEPETERSEN, R. and GUAN, Y. (2023). Semiparametric goodness-of-fit test for clustered point processes with a shape-constrained pair correlation function. *J. Amer. Statist. Assoc.* **118** 2072–2087. MR4646627 https://doi.org/10.1080/01621459.2022.2029456

XU, J. and GRIFFITHS, E. (2017). Shooting on the street: Measuring the spatial influence of physical features on gun violence in a bounded street network. *J. Quant. Criminol.* **33** 237–253.

YUAN, B., LI, H., BERTOZZI, A. L., BRANTINGHAM, P. J. and PORTER, M. A. (2019). Multivariate spatiotemporal Hawkes processes and network reconstruction. *SIAM J. Math. Data Sci.* **1** 356–382. MR3975150 https://doi.org/10.1137/18M1226993

YUAN, N. J., ZHENG, Y., XIE, X., WANG, Y., ZHENG, K. and XIONG, H. (2014). Discovering urban functional zones using latent activity trajectories. *IEEE Trans. Knowl. Data Eng.* **27** 712–725.

ZHANG, Q., LIPANI, A., KIRNAP, O. and YILMAZ, E. (2020). Self-attentive Hawkes process. In *International Conference on Machine Learning* 11183–11193. PMLR.

ZHU, S., LI, S., PENG, Z. and XIE, Y. (2021a). Imitation learning of neural spatio-temporal point processes. *IEEE Trans. Knowl. Data Eng.* **34** 5391–5402.

ZHU, S., WANG, H., DONG, Z., CHENG, X. and XIE, Y. (2021b). Neural spectral marked point processes. In *International Conference on Learning Representations*.

ZHU, S. and XIE, Y. (2022). Spatiotemporal-textual point processes for crime linkage detection. *Ann. Appl. Stat.* **16** 1151–1170. MR4438828 https://doi.org/10.1214/21-aoas1538

ZHU, S., ZHANG, M., DING, R. and XIE, Y. (2021c). Deep Fourier kernel for self-attentive point processes. In *International Conference on Artificial Intelligence and Statistics* 856–864. PMLR.

ZHUANG, J. and MATEU, J. (2019). A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data. *J. Roy. Statist. Soc., Ser. A, Statist. Soc.* **182** 919–942. MR3955503 https://doi.org/10.1111/rssa.12429

ZIPKIN, J. R., SHORT, M. B. and BERTOZZI, A. L. (2014). Cops on the dots in a mathematical model of urban crime and police response. *Discrete Contin. Dyn. Syst. Ser. B* **19** 1479–1506. MR3199788 https://doi.org/10.3934/dcdsb.2014.19.1479

ZUO, S., JIANG, H., LI, Z., ZHAO, T. and ZHA, H. (2020). Transformer Hawkes process. In *International Conference on Machine Learning* 11692–11702. PMLR.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1671–1690
https://doi.org/10.1214/26-AOAS2171

DYNAMIC COUNT MODELS WITH FLEXIBLE INNOVATION PROCESSES FOR IRREGULAR MARITIME MIGRATION

BY GREGOR ZENS[1,a] AND JAKUB BIJAK[2,b]

[1]*Population and Just Societies Program, International Institute for Applied Systems Analysis (IIASA),* [a]*zens@iiasa.ac.at*
[2]*Leverhulme Centre for Demographic Science, Nuffield Department of Population Health and Reuben College, University of Oxford,* [b]*jakub.bijak@demography.ox.ac.uk*

Motivated by the challenge of analyzing the dynamics of weekly sea border crossings in the Mediterranean (2015–2025) and the English Channel (2018–2025), we develop a Bayesian dynamic framework for modeling heteroskedastic count time series. Building on theoretical considerations and empirical stylized facts, our approach utilizes a Poisson random walk model that allows for heavy-tailed innovations or stochastic volatility dynamics, while incorporating an explicit mechanism to separate structural from sampling zeros. Posterior inference is carried out via a straightforward Markov chain Monte Carlo algorithm. Applying this methodology to Mediterranean and English Channel data, we compare alternative model specifications through a comprehensive out-of-sample forecasting exercise. Using log predictive scores and empirical coverage at predictive quantiles to evaluate each model, we find strong evidence for stochastic volatility in migration innovations. These models deliver the strongest out-of-sample forecasts with empirical coverage close to nominal levels up to the 99th percentile. Our framework can be used to develop risk indicators with direct policy implications for improving governance and preparedness for migration surges. More broadly, the methodology extends to other zero-inflated nonstationary count time series applications, including epidemiological surveillance and public safety incident monitoring.

REFERENCES

AKTEKIN, T., SOYER, R. and ZHANG, D. (2026). Bayesian forecasting of zero-inflated time-series of counts. *Int. J. Forecast*.

BARRA, I., BOROWSKA, A. and KOOPMAN, S. J. (2018). Bayesian dynamic modeling of high-frequency integer price changes. *J. Financ. Econom.* **16** 384–424.

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

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

BHATIYA, A. and KADAM, S. (2025). Small Boats, Big Impacts: the Ripple Effects of Irregular Migration Technical Report, Competitive Advantage in the Global Economy (CAGE).

BIJAK, J. (2010). *Forecasting International Migration in Europe*: *A Bayesian View*. Springer, Berlin.

BIJAK, J., ed. (2024). *From Uncertainty to Policy*: *A Guide to Migration Scenarios* Edward Elgar, Cheltenham Glos.

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

BOSCO, C., MINORA, U., DE RIGO, D., PINGSDORF, J. and CORTINOVIS, R. (2025). Supporting migration policies with forecasts: Illegal border crossings in Europe through a mixed approach. arXiv preprint. Available at arXiv:2512.10633.

BOSCO, C., MINORA, U., ROSIŃSKA, A., TEOBALDELLI, M. and BELMONTE, M. (2024). A machine learning architecture to forecast irregular border crossings and asylum requests for policy support in Europe: A case study. *Data Policy* **6** e81.

Key words and phrases. Zero-inflated Poisson, Bayesian dynamic models, migration time series, distributional forecasting, heavy-tailed innovations.

CAMARENA, K. R., CLAUDY, S., WANG, J. and WRIGHT, A. L. (2020). Political and environmental risks influence migration and human smuggling across the Mediterranean Sea. *PLoS ONE* **15** e0236646.

CARAMMIA, M., IACUS, S. M. and WILKIN, T. (2022). Forecasting asylum-related migration flows with machine learning and data at scale. *Sci. Rep.* **12** 1457.

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

CZAIKA, M. and HOBOLTH, M. (2016). Do restrictive asylum and visa policies increase irregular migration into Europe? *Eur. Union Polit.* **17** 345–365. https://doi.org/10.1177/1465116516633299

DAVIS, R. A., HOLAN, S. H., LUND, R. and RAVISHANKER, N., eds. (2016). *Handbook of Discrete-Valued Time Series. Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR3642975

DEIANA, C., MAHESHRI, V. and MASTROBUONI, G. (2024). Migrants at sea: Unintended consequences of search and rescue operations. *Amer. Econ. J., Econ. Policy* **16** 335–365.

EC (2020). Commission Recommendation (EU) 2020/1366 of 23 September 2020 on an EU mechanism for preparedness and management of crises related to migration (Migration Preparedness and Crisis Blueprint). *OJ L* **317** 26–38. 1.10.2020. Available at https://eur-lex.europa.eu/eli/reco/2020/1366/oj/eng.

FRIEBEL, G., MANCHIN, M., MENDOLA, M. and PRAROLO, G. (2024). International migration and illegal costs: Evidence from Africa-to-Europe smuggling routes. *J. Int. Econ.* **148** 103878.

FRONTEX (2024). Migratory Routes. Available at https://www.frontex.europa.eu/what-we-do/monitoring-and-risk-analysis/migratory-routes/migratory-routes/.

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

GEORGIOU, H. V. (2016). Identification of refugee influx patterns in Greece via model-theoretic analysis of daily arrivals. arXiv preprint. Available at arXiv:1605.02784.

HM GOVERNMENT (2025). National Risk Register: 2025 Edition. Available at https://bit.ly/uk-nrr-2025.

HOFFMANN PHAM, K. and KOMIYAMA, J. (2024). Strategic choices of migrants and smugglers in the central Mediterranean Sea. *PLoS ONE* **19** e0300553.

IPCC (2012). Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. Special Report of the Intergovernmental Panel on Climate Change. Cambridge Univ. Press, Cambridge, MA. Available at https://www.cambridge.org/9781107607804.

JUNG, R. C., KUKUK, M. and LIESENFELD, R. (2006). Time series of count data: Modeling, estimation and diagnostics. *Comput. Statist. Data Anal.* **51** 2350–2364. MR2307505 https://doi.org/10.1016/j.csda.2006.08.001

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

KIM, S. and ALBERT, P. S. (2018). Latent variable Poisson models for assessing the regularity of circadian patterns over time. *J. Amer. Statist. Assoc.* **113** 992–1002. MR3862334 https://doi.org/10.1080/01621459.2017.1379402

KIM, S., SHEPHARD, N. and CHIB, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Rev. Econ. Stud.* **65** 361–393.

KOWAL, D. R., MATTESON, D. S. and RUPPERT, D. (2019). Dynamic shrinkage processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 781–804. MR3997101 https://doi.org/10.1111/rssb.12325

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

MIGRATION OBSERVATORY (2025). People crossing the English Channel in small boats. Available at https://migrationobservatory.ox.ac.uk/resources/briefings/people-crossing-the-english-channel-in-small-boats.

MILLS, C., IRONS, N. J., TSUI, J. L. H., SPARROW, S., CARVALHO, L. M., KUCHARSKI, A. J., RATMANN, O., LAMBERT, B., DONNELLY, C. A. et al. (2025). From metric to action: An evaluation framework to translate infectious disease forecasts into policy decisions. medRxiv. https://doi.org/10.1101/2025.07.20.25331802

MUTISO, F., PEARCE, J. L., BENJAMIN-NEELON, S. E., MUELLER, N. T., LI, H. and NEELON, B. (2022). Bayesian negative binomial regression with spatially varying dispersion: Modeling COVID-19 incidence in Georgia. *Spat. Stat.* **52** Paper No. 100703, 19. MR4496717 https://doi.org/10.1016/j.spasta.2022.100703

NEELON, B. (2019). Bayesian zero-inflated negative binomial regression based on Pólya-Gamma mixtures. *Bayesian Anal.* **14** 829–855. MR3960773 https://doi.org/10.1214/18-BA1132

ROBERTS, G. O. and ROSENTHAL, J. S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist.* **18** 349–367. MR2749836 https://doi.org/10.1198/jcgs.2009.06134

RODRÍGUEZ SÁNCHEZ, A., WUCHERPFENNIG, J., RISCHKE, R. and IACUS, S. M. (2023). Search-and-rescue in the central Mediterranean route does not induce migration: Predictive modeling to answer causal queries in migration research. *Sci. Rep.* **13** 11014.

SCHAFER, T. L. J. and MATTESON, D. S. (2025). Locally adaptive shrinkage priors for trends and breaks in count time series. *Technometrics* **67** 157–167. MR4866660 https://doi.org/10.1080/00401706.2024.2407316

WEST, M., HARRISON, P. J. and MIGON, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *J. Amer. Statist. Assoc.* **80** 73–83. With discussion. MR0786598 https://doi.org/10.1080/01621459.1985.10477131

WOOD, R. M. (2025). Factors affecting illegal migration to the UK by small-boat crossing of the English channel: Descriptive, inferential and predictive analyses. *Int. Migr. Rev.* https://doi.org/10.1177/0197918325139400

ZAMBIASI, D. and ALBAROSA, E. (2025). Externalizing rescue operations at sea: The migration deal between Italy and Libya. *J. Econ. Geogr.* **25** 41–58.

ZENS, G. and BIJAK, J. (2026). Supplement to "Dynamic count models with flexible innovation processes for irregular maritime migration." https://doi.org/10.1214/26-AOAS2171SUPP

ZENS, G. and THALHEIMER, L. (2025). The short-term dynamics of conflict-driven displacement: Bayesian modeling of disaggregated data from Somalia. *Ann. Appl. Stat.* **19** 286–301. MR4888111 https://doi.org/10.1214/24-aoas1959

ZHANG, L., ZHOU, M., YU, K. and TIAN, M. (2026). A spatiotemporal marginalized zero-inflated Conway–Maxwell–Poisson regression model: Application to international population outmigration within Asia. *J. Roy. Statist. Soc., Ser. A, Statist. Soc.* https://doi.org/10.1093/jrsssa/qnag009

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1691–1718
https://doi.org/10.1214/26-AOAS2163

PERTURBATION-ROBUST PREDICTIVE MODELING OF SOCIAL EFFECTS BY NETWORK SUBSPACE GENERALIZED LINEAR MODELS

BY JIANXIANG WANG[1,a], CAN M. LE[2,b] AND TIANXI LI[3,c]

[1]*Department of Statistics, Rutgers University—New Brunswick,* [a]*jw1881@scarletmail.rutgers.edu*
[2]*Department of Statistics, University of California, Davis,* [b]*canle@ucdavis.edu*
[3]*School of Statistics, University of Minnesota, Twin Cities,* [c]*tianxili@umn.edu*

Network-linked data, in which multivariate observations are interconnected by a network, are becoming increasingly prevalent in fields such as sociology and biology. These data often exhibit inherent noise and complex relational structures, complicating conventional modeling and statistical inference. Motivated by empirical challenges in analyzing such datasets, this paper introduces a family of network subspace generalized linear models designed for analyzing noisy, network-linked data. We propose a model inference method based on subspace-constrained maximum likelihood that emphasizes flexibility in capturing network effects and provides an inference framework that is robust under network perturbations. We establish the asymptotic distributions of the estimators under network perturbations, demonstrating the method's accuracy through extensive simulations involving random network models and deep-learning-based embedding algorithms. The proposed methodology is applied to a comprehensive analysis of a large-scale study on school conflicts, where it identifies significant social effects, offering meaningful and interpretable insights into student behavior.

REFERENCES

ARMILLOTTA, M. and FOKIANOS, K. (2023). Nonlinear network autoregression. *Ann. Statist.* **51** 2526–2552. MR4682707 https://doi.org/10.1214/23-aos2345

ATHREYA, A., FISHKIND, D. E., TANG, M., PRIEBE, C. E., PARK, Y., VOGELSTEIN, J. T., LEVIN, K., LYZINSKI, V., QIN, Y. et al. (2018). Statistical inference on random dot product graphs: A survey. *J. Mach. Learn. Res.* **18** Paper No. 226, 92. MR3827114

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

CHANG, J. H. and PAUL, S. (2024). Embedding Network Autoregression for time series analysis and causal peer effect inference. arXiv preprint. Available at arXiv:2406.05944.

CHATTERJEE, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43** 177–214. MR3285604 https://doi.org/10.1214/14-AOS1272

CUADRA, L., SALCEDO-SANZ, S., DEL SER, J., JIMÉNEZ-FERNÁNDEZ, S. and GEEM, Z. W. (2015). A critical review of robustness in power grids using complex networks concepts. *Energies* **8** 9211–9265.

FANG, G., XU, G., XU, H., ZHU, X. and GUAN, Y. (2024). Group network Hawkes process. *J. Amer. Statist. Assoc.* **119** 2328–2344. MR4797944 https://doi.org/10.1080/01621459.2023.2257889

GAO, Q.-B., LIN, J.-G., ZHU, C.-H. and WU, Y.-H. (2012). Asymptotic properties of maximum quasi-likelihood estimators in generalized linear models with adaptive designs. *Statistics* **46** 833–846. MR2989781 https://doi.org/10.1080/02331888.2010.543465

GREEN, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. Roy. Statist. Soc. Ser. B, Methodol.* **46** 149–192. With discussion. MR0781879

GROVER, A. and LESKOVEC, J. (2016). Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

HAN, X., YANG, Q. and FAN, Y. (2023). Universal rank inference via residual subsampling with application to large networks. *Ann. Statist.* **51** 1109–1133. MR4630942 https://doi.org/10.1214/23-aos2282

Key words and phrases. Generalized linear model, network-linked data, network perturbation, network embedding.

HARRIS, K. M. (2009). The National Longitudinal Study of Adolescent to Adult Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002; Wave IV, 2007–009 [machine-readable data file and documentation]. Carolina Population Center, Univ. North Carolina at Chapel Hill.

HAYES, A., FREDRICKSON, M. M. and LEVIN, K. (2022). Estimating network-mediated causal effects via spectral embeddings. arXiv preprint. Available at arXiv:2212.12041.

HE, Y., SUN, J., TIAN, Y., YING, Z. and FENG, Y. (2025). Semiparametric modeling and analysis for longitudinal network data. *Ann. Statist.* **53** 1406–1430. MR4959800 https://doi.org/10.1214/25-AOS2506

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

HOLME, P. (2015). Modern temporal network theory: A colloquium. *Eur. Phys. J. B* **88** 1–30.

KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83** 016107, 10. MR2788206 https://doi.org/10.1103/PhysRevE.83.016107

KIPF, T. N. and WELLING, M. (2016). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

LE, C. M. and LEVINA, E. (2022). Estimating the number of communities by spectral methods. *Electron. J. Stat.* **16** 3315–3342. MR4422967 https://doi.org/10.1214/21-ejs1971

LE, C. M. and LI, T. (2022). Linear regression and its inference on noisy network-linked data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 1851–1885. MR4515560

LE GAT, Y. (2014). Extending the Yule process to model recurrent pipe failures in water supply networks. *Urban Water J.* **11** 617–630.

LEE, S. Y. (2019). Document vectorization method using network information of words. *PLoS ONE* **14** e0219389.

LEE, Y. and OGBURN, E. L. (2021). Network dependence can lead to spurious associations and invalid inference. *J. Amer. Statist. Assoc.* **116** 1060–1074. MR4309250 https://doi.org/10.1080/01621459.2020.1782219

LI, T. and LE, C. M. (2024). Network estimation by mixing: Adaptivity and more. *J. Amer. Statist. Assoc.* **119** 2190–2205. MR4797933 https://doi.org/10.1080/01621459.2023.2252137

LI, T., LEVINA, E. and ZHU, J. (2019). Prediction models for network-linked data. *Ann. Appl. Stat.* **13** 132–164. MR3937424 https://doi.org/10.1214/18-AOAS1205

LI, T., LEVINA, E. and ZHU, J. (2020). Network cross-validation by edge sampling. *Biometrika* **107** 257–276. MR4108931 https://doi.org/10.1093/biomet/asaa006

LI, T., LEVINA, E., ZHU, J. and LE, C. M. (2023). randnet: Random Network Model Estimation, Selection and Parameter Tuning R package version 0.7.

LIU, X. and HUANG, K.-W. (2025). Controlling homophily in social network regression analysis by machine learning. *INFORMS J. Comput.* **37** 684–702. MR4923945

LUNDE, R., LEVINA, E. and ZHU, J. (2025). Conformal prediction for network-assisted regression. *J. Amer. Statist. Assoc.* **120** 1633–1644. MR4973885 https://doi.org/10.1080/01621459.2025.2506198

MA, Z., MA, Z. and YUAN, H. (2020). Universal latent space model fitting for large networks with edge covariates. *J. Mach. Learn. Res.* **21** Paper No. 4, 67. MR4071187

MAO, X., CHAKRABARTI, D. and SARKAR, P. (2021). Consistent nonparametric methods for network assisted covariate estimation. In *International Conference on Machine Learning* 7435–7446. PMLR.

MCCULLAGH, P. (2019). *Generalized Linear Models*. Routledge, London.

MICHELL, L. and PEARSON, M. (2000). Smoke rings: Social network analysis of friendship groups, smoking and drug-taking. *Drugs Educ. Prev. Policy* **7** 21–37.

MICHELL, L. and WEST, P. (1996). Peer pressure to smoke: The meaning depends on the method. *Health Educ. Res.* **11** 39–49.

MUKHERJEE, S., NIU, Z., HALDER, S., BHATTACHARYA, B. B. and MICHAILIDIS, G. (2021). High dimensional logistic regression under network dependence. arXiv preprint. Available at arXiv:2110.03200.

NEWMAN, M. E. J. (2003). Mixing patterns in networks. *Phys. Rev. E* **67** 026126, 13. MR1975193 https://doi.org/10.1103/PhysRevE.67.026126

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

ÖZGÜR, A., VU, T., ERKAN, G. and RADEV, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics* **24** i277–i285.

PALUCK, E. L., SHEPHERD, H. and ARONOW, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proc. Natl. Acad. Sci. USA* **113** 566–571.

PEROZZI, B., AL-RFOU, R. and SKIENA, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 701–710.

PHAN, T. Q. and AIROLDI, E. M. (2015). A natural experiment of social network formation and dynamics. *Proc. Natl. Acad. Sci. USA* **112** 6595–6600.

POZEK, M., SIKIC, L., AFRIC, P., KURDIJA, A. S., VLADIMIR, K., DELAC, G. and SILIC, M. (2019). Performance of common classifiers on node2vec network representations. In 2019 42*nd International Convention on Information and Communication Technology, Electronics and Microelectronics* (*MIPRO*) 925–930. IEEE Press, New York.

PRANATHI, K. S. and PRATHIBHAMOL, C. (2021). Node classification through graph embedding techniques. In 2021 4*th Biennial International Conference on Nascent Technologies in Engineering* (*ICNTE*) 1–4. IEEE Press, New York.

ROZEMBERCZKI, B., KISS, O. and SARKAR, R. (2020). Karate club: An API oriented open-source python framework for unsupervised learning on graphs. In *Proceedings of the* 29*th. ACM International Conference on Information and Knowledge Management* (*CIKM* '20) 3125–3132. ACM, New York.

ROZEMBERCZKI, B. and SARKAR, R. (2018). Fast sequence-based embedding with diffusion graphs. In *Complex Networks IX*: *Proceedings of the* 9*th Conference on Complex Networks CompleNet* 2018 **9** 99–107. Springer, Berlin.

RUBIN-DELANCHY, P., CAPE, J., TANG, M. and PRIEBE, C. E. (2022). A statistical interpretation of spectral embedding: The generalised random dot product graph. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 1446–1473. MR4494166 https://doi.org/10.1111/rssb.12509

SCARSELLI, F., GORI, M., TSOI, A. C., HAGENBUCHNER, M. and MONFARDINI, G. (2008). The graph neural network model. *IEEE Trans. Neural Netw.* **20** 61–80.

SINCLAIR, B., MCCONNELL, M. and GREEN, D. P. (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *Amer. J. Polit. Sci.* **56** 1055–1069.

SIT, T., YING, Z. and YU, Y. (2021). Event history analysis of dynamic networks. *Biometrika* **108** 223–230. MR4226200 https://doi.org/10.1093/biomet/asaa045

SU, L., LU, W., SONG, R. and HUANG, D. (2020). Testing and estimation of social network dependence with time to event data. *J. Amer. Statist. Assoc.* **115** 570–582. MR4107658 https://doi.org/10.1080/01621459.2019.1617153

VAN DEN BOS, W., CRONE, E. A., MEUWESE, R. and GÜROĞLU, B. (2018). Social network cohesion in school classes promotes prosocial behavior. *PLoS ONE* **13** e0194656.

WANG, J., LE, C. M. and LI, T. (2026). Supplement to "Perturbation-robust predictive modeling of social effects by network subspace generalized linear models." https://doi.org/10.1214/26-AOAS2163SUPPA, https://doi.org/10.1214/26-AOAS2163SUPPB

WU, W. and LENG, C. (2023). A random graph-based autoregressive model for networked time series. arXiv preprint. Available at arXiv:2309.08488.

YIN, C., ZHAO, L. and WEI, C. (2006). Asymptotic normality and strong consistency of maximum quasi-likelihood estimates in generalized linear models. *Sci. China Ser. A* **49** 145–157. MR2223705 https://doi.org/10.1007/s11425-004-5169-x

YU, H., BRAUN, P., YILDIRIM, M. A., LEMMENS, I., VENKATESAN, K., SAHALIE, J., HIROZANE-KISHIKAWA, T., GEBREAB, F., LI, N. et al. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science* **322** 104–110.

ZENG, X., LIU, L., LÜ, L. and ZOU, Q. (2018). Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* **34** 2425–2432.

ZHANG, X., PAN, R., GUAN, G., ZHU, X. and WANG, H. (2020). Logistic regression with network structure. *Statist. Sinica* **30** 673–693. MR4214157

ZHANG, Y., LEVINA, E. and ZHU, J. (2016). Community detection in networks with node features. *Electron. J. Stat.* **10** 3153–3178. MR3571965 https://doi.org/10.1214/16-EJS1206

ZHANG, Y., LEVINA, E. and ZHU, J. (2017). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika* **104** 771–783. MR3737303 https://doi.org/10.1093/biomet/asx042

ZHANG, Y. and TANG, M. (2023). A theoretical analysis of DeepWalk and Node2vec for exact recovery of community structures in stochastic blockmodels. *IEEE Trans. Pattern Anal. Mach. Intell.*

ZHU, X., PAN, R., LI, G., LIU, Y. and WANG, H. (2017). Network vector autoregression. *Ann. Statist.* **45** 1096–1123. MR3662449 https://doi.org/10.1214/16-AOS1476

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1719–1735
https://doi.org/10.1214/26-AOAS2162

MEASURING PUBLIC OPINION: "THE WASSERSTEIN BIPOLARIZATION INDEX," WITH APPLICATION TO CROSS-NATIONAL ATTITUDES TOWARD MANDATORY VACCINATION FOR COVID-19

BY HANE LEE[a] AND MICHAEL E. SOBEL[b]

Department of Statistics, Columbia University, [a]*hane.lee@columbia.edu,* [b]*mes105@columbia.edu*

The extent to which the American public is politically polarized is of great interest in the lay and academic communities. To study opinion polarization, political scientists and public opinion researchers examine the distribution of respondents on survey items, using visual comparison of histograms and/or measures such as variances and bimodality coefficients. We prove these measures fail to align with prevailing conceptualizations of polarization put forth in the literature. To remedy this situation, we specify several properties a measure of polarization consistent with these conceptualizations should possess: in particular, it should increase as a distribution spreads away from a center toward the poles and/or as clustering below or above this center increases. We then propose a p-Wasserstein bipolarization index that satisfies these properties and measures the distance between the distribution of an item and a most polarized distribution with all mass concentrated on the lower and upper endpoints of the scale, using the index to examine bipolarization in attitudes toward governmental COVID-19 vaccine mandates across 11 countries: the U.S. and U.K. are most polarized, China, France, and India the least polarized, with Spain, Colombia, Italy, Brazil, Australia, and Canada occupying an intermediate position.

REFERENCES

ALLISON, R. A. and FOSTER, J. E. (2004). Measuring health inequality using qualitative data. *J. Health Econ*. **23** 505–524. https://doi.org/10.1016/j.jhealeco.2003.10.006

ANDERSON, G. (2004). Making inferences about the polarization, welfare and poverty of nations: A study of 101 countries 1970–1995. *J. Appl. Econometrics* **19** 537–550. https://doi.org/10.1002/jae.750

APOUEY, B. (2007). Measuring health polarization with self-assessed health data. *Health Econ*. **16** 875–894. https://doi.org/10.1002/hec.1284

APOUEY, B. and SILBER, J. (2013). Inequality and bi-polarization in socioeconomic status and health: Ordinal approaches. In *Research on Economic Inequality*: *Health and Inequality* **21** 77–109. Emerald Group Publishing Limited, Leeds, UK. https://doi.org/10.1108/S1049-2585(2013)0000021005

BAUER, P. (2019). Conceptualizing and measuring polarization: a review. SocArXiv. https://doi.org/10.31235/osf.io/e5vp8

BAUER, P. C., BARBERÁ, P., ACKERMANN, K. and VENETZ, A. (2017). Is the left-right scale a valid measure of ideology?: Individual-level variation in associations with "left" and "right" and left-right self-placement. *Polit. Behav.* **39** 553–583. https://doi.org/10.1007/s11109-016-9368-2

BOLSEN, T. and PALM, R. (2022). Politicization and COVID-19 vaccine resistance in the U.S. In *Progress in Molecular Biology and Translational Science* (T. Bolsen and R. Palm, eds.) **188** 81–100. Academic Press, Cambridge, MA. https://doi.org/10.1016/bs.pmbts.2021.10.002

COWELL, F. A. (2000). Measurement of inequality. In *Handbook of Income Distribution* (A. B. Atkinson and F. Bourguignon, eds.) **1** 87–166. Elsevier, Amsterdam, Netherlands. Chapter 2. https://doi.org/10.1016/S1574-0056(00)80005-6

DIMAGGIO, P., EVANS, J. and BRYSON, B. (1996). Have American's social attitudes become more polarized? *Amer. J. Sociol*. **102** 690–755. https://doi.org/10.1086/230995

DOWNS, A. (1957). *An Economic Theory of Democracy*. Harper & Brothers, New York, NY.

DRUCKMAN, J. N. and LEEPER, T. J. (2012). Is public opinion stable? Resolving the micro/macro disconnect in studies of public opinion. *Daedalus* **141** 50–68. https://doi.org/10.1162/DAED_a_00173

Key words and phrases. Polarization, Wasserstein distance, public opinion, COVID-19 vaccination mandates.

DUCH, R., ROOPE, L. S. J., VIOLATO, M., BECERRA, M. F., ROBINSON, T. S., BONNEFON, J.-F., FRIEDMAN, J., LOEWEN, P. J., MAMIDI, P. et al. (2021). Citizens from 13 countries share similar preferences for COVID-19 vaccine allocation priorities. *Proc. Natl. Acad. Sci. USA* **118** e2026382118. https://doi.org/10.1073/pnas.2026382118

DUCLOS, J.-Y., ESTEBAN, J. and RAY, D. (2004). Polarization: Concepts, measurement, estimation. *Econometrica* **72** 1737–1772. MR2095531 https://doi.org/10.1111/j.1468-0262.2004.00552.x

ENDERS, A. M. (2021). Issues versus affect: How do elite and mass polarization compare? *J. Polit.* **83** 1872–1877. https://doi.org/10.1086/715059

ESTEBAN, J., GRADÍN, C. and RAY, D. (2007). An extension of a measure of polarization, with an application to the income distribution of five OECD countries. *J. Econ. Inequal.* **5** 1–19. https://doi.org/10.1007/s10888-006-9032-x

ESTEBAN, J. and RAY, D. (1994). On the measurement of polarization. *Econometrica* **62** 819–851.

ESTEBAN, J. and RAY, D. (2012). Comparing polarization measures. In *The Oxford Handbook of the Economics of Peace and Conflict*. https://doi.org/10.1093/oxfordhb/9780195392777.013.0007

FIORINA, M. P. (2017). *Unstable Majorities*: *Polarization*, *Party Sorting*, *and Political Stalemate*. Hoover Institution Press, Stanford, CA.

FIORINA, M. P. and ABRAMS, S. J. (2008). Political polarization in the American public. *Annu. Rev. Pol. Sci.* **11** 563–588. https://doi.org/10.1146/annurev.polisci.11.053106.153836

FOSTER, J. and WOLFSON, M. (2010). Polarization and the decline of the middle class in Canada and the USA. *J. Econ. Inequal.* **8** 247–273. https://doi.org/10.1007/s10888-009-9122-7

GILENS, M. and PAGE, B. I. (2014). Testing theories of American politics: Elites, interest groups, and average citizens. *Perspect. Polit.* **12** 564–581. https://doi.org/10.1017/S1537592714001595

GRAMACHO, W. G. and TURGEON, M. (2021). When politics collides with public health: COVID-19 vaccine country of origin and vaccination acceptance in Brazil. *Vaccine* **39** 2608–2612. https://doi.org/10.1016/j.vaccine.2021.03.080

HAN, X., WANG, J., ZHANG, M. and WANG, X. (2020). Using social media to mine and analyze public opinion related to COVID-19 in China. *Int. J. Environ. Res. Public Health* **17** 2788. https://doi.org/10.3390/ijerph17082788

HILL, S. J. and TAUSANOVITCH, C. (2015). A disconnect in representation? Comparison of trends in congressional and public polarization. *J. Polit.* **77** 1058–1075. https://doi.org/10.1086/682398

KIM, H. and FORDING, R. C. (2003). Voter ideology in western democracies: An update. *Eur. J. Polit. Res.* **42** 95–105. https://doi.org/10.1111/1475-6765.00076

KOBUS, M. (2015). Polarization measurement for ordinal data. *J. Econ. Inequal.* **13** 275–297. https://doi.org/10.1007/s10888-014-9282-y

KOBUS, M., KAPERA, M. and MAASOUMI, E. (2024). Gap in many dimensions: Application to gender. *Labour Econ.* **89** 102582. https://doi.org/10.1016/j.labeco.2024.102582

LEE, H. and SOBEL, M. E. (2026). Supplement to "Measuring public opinion: "the Wasserstein Bipolarization Index," with application to cross-national attitudes toward mandatory vaccination for COVID-19." https://doi.org/10.1214/26-AOAS2162SUPP

LELKES, Y. (2016). Mass polarization: Manifestations and measurements. *Public Opin. Q.* **80** 392–410. https://doi.org/10.1093/poq/nfw005

MACHINA, M. J. and PRATT, J. W. (1997). Increasing risk: Some direct constructions. *J. Risk Uncertain.* **14** 103–127. https://doi.org/10.1023/A:1007719626543

MOUW, T. and SOBEL, M. E. (2001). Culture wars and opinion polarization: The case of abortion. *Amer. J. Sociol.* **106** 913–943. https://doi.org/10.1086/320294

PEYRÉ, G. and CUTURI, M. (2019). Computational optimal transport: With applications to data science. *Found. Trends Mach. Learn.* **11** 355–607. https://doi.org/10.1561/2200000073

ROTHSCHILD, M. and STIGLITZ, J. E. (1970). Increasing risk. I. A definition. *J. Econom. Theory* **2** 225–243. MR0503565 https://doi.org/10.1016/0022-0531(70)90038-4

SARKAR, S. and SANTRA, S. (2020). Extending the approaches to polarization ordering of ordinal variables. *J. Econ. Inequal.* **18** 421–440. https://doi.org/10.1007/s10888-020-09442-x

SOMMERFELD, M. (2017). Wasserstein distance on finite spaces: Statistical inference and algorithms. PhD thesis, Universität Göttingen.

THORPE, M. (2018). Introduction to optimal transport. Notes of Course at Univ. Cambridge.

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

WANG, Y.-Q. and TSUI, K.-Y. (2000). Polarization orderings and new classes of polarization indices. *J. Public Econ. Theory* **2** 349–363. https://doi.org/10.1111/1097-3923.00042

ZALLER, J. R. (1992). *The Nature and Origins of Mass Opinion*. Cambridge Univ. Press, Cambridge, UK.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1736–1765
https://doi.org/10.1214/26-AOAS2183

A STRUCTURED ESTIMATOR FOR LARGE COVARIANCE MATRICES IN THE PRESENCE OF PAIRWISE AND SPATIAL COVARIATES

BY MARTIN METODIEV[1,a], MARIE PERROT-DOCKÈS[2,c], SARAH OUADAH[3,d], BAILEY K. FOSDICK[4,f], STÉPHANE ROBIN[3,e], PIERRE LATOUCHE[1,5,b] AND ADRIAN E. RAFTERY[6,g]

[1]*Laboratoire de Mathématiques Blaise Pascal, Université Clermont Auvergne, CNRS,* [a]*martin.metodiev@doctorant.uca.fr,* [b]*Pierre.LATOUCHE@uca.fr*

[2]*Université Paris Cité, CNRS, MAP5,* [c]*marie.perrot-dockees@u-paris.fr*

[3]*Sorbonne Université, Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation (LPSM),* [d]*sarah.ouadah@sorbonne-universite.fr,* [e]*stephane.robin@sorbonne-universite.fr*

[4]*Department of Biostatistics & Informatics, Colorado School of Public Health,* [f]*bailey.fosdick@cuanschutz.edu*

[5]*Institut Universitaire de France (IUF)*

[6]*Department of Statistics, University of Washington,* [g]*raftery@uw.edu*

We consider the problem of estimating a high-dimensional covariance matrix from a small number of observations when covariates on pairs of variables are available and the variables can have spatial structure. This is motivated by the problem arising in demography of estimating the covariance matrix of the total fertility rate (TFR) of 195 different countries when only 11 observations are available. We construct an estimator for high-dimensional covariance matrices by exploiting information about pairwise covariates, such as whether pairs of variables belong to the same cluster, or spatial structure of the variables, and interactions between the covariates. We reformulate the problem in terms of a mixed effects model. This requires the estimation of only a small number of parameters, which are easy to interpret and which can be selected using standard procedures. The estimator is consistent under general conditions, and asymptotically normal. It works if the mean and variance structure of the data is already specified or if some of the data are missing. Using simulations, we assess its performance under our model assumptions as well as under model misspecification. We find that it outperforms several popular alternatives. We apply it to the TFR dataset and draw some conclusions.

REFERENCES

AGUILAR, O. and WEST, M. (2000). Bayesian dynamic factor models and portfolio allocation. *J. Bus. Econom. Statist.* **18** 338–357.

ALKEMA, L., RAFTERY, A. E., GERLAND, P., CLARK, S. J., PELLETIER, F., BUETTNER, T. and HEILIG, G. K. (2011a). Probabilistic projections of the total fertility rate for all countries. *Demography* **48** 815–839.

ALKEMA, L., RAFTERY, A. E., GERLAND, P., CLARK, S. J., PELLETIER, F., BUETTNER, T. and HEILIG, G. K. (2011b). Probabilistic projections of the total fertility rate for all countries, Online Resource 1. *Demography* **48** 815–839.

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

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

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

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

BATES, D., MAECHLER, M. and JAGAN, M. (2024). Matrix: Sparse and dense matrix classes and methods. R package version 1.7-0.

Key words and phrases. Large covariance matrix estimation, pairwise covariates, spatial effects.

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., 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

BONAT, W. H. and JØRGENSEN, B. (2016). Multivariate covariance generalized linear models. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **65** 649–675. MR3564998 https://doi.org/10.1111/rssc.12145

BROYDEN, C. G. (1970). The convergence of a class of double-rank minimization algorithms I. General considerations. *IMA J. Appl. Math.* **6** 76–90. https://doi.org/10.1093/imamat/6.1.76

BURG, J. P., LUENBERGER, D. G. and WENGER, D. L. (1983). Estimation of structured covariance matrices. *Proc. IEEE* **70** 963–974.

CHENG, S. H. and HIGHAM, N. J. (1998). A modified Cholesky algorithm based on a symmetric indefinite factorization. *SIAM J. Matrix Anal. Appl.* **19** 1097–1110. MR1636528 https://doi.org/10.1137/S0895479896302898

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

CHRISTENSEN, W. F. and AMEMIYA, Y. (2003). Modeling and prediction for multivariate spatial factor analysis. *J. Statist. Plann. Inference* **115** 543–564. MR1985883 https://doi.org/10.1016/S0378-3758(02)00173-8

COX, D. R. and REID, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91** 729–737. MR2090633 https://doi.org/10.1093/biomet/91.3.729

KAROLYI, G. A. (1993). A Bayesian approach to modeling stock return volatility for option valuation. *J. Financ. Quant. Anal.* **28** 579–594.

DE BEER, J. (2000). *Dealing with Uncertainty in Population Forecasting*. Statistics Netherlands, Department of Population, Voorburg.

DE FREITAS, L. A. C., CARLOS, LDO., CAMPOS, A. C. L. and BONAT, W. H. (2022). Hypothesis tests for multiple responses regression: Effect of probiotics on addiction and binge eating disorder. arXiv e-prints, arXiv:2208.00027.

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

FAN, J., LIAO, Y. and LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econom. J.* **19** C1–C32. MR3501529 https://doi.org/10.1111/ectj.12061

FLETCHER, R. and REEVES, C. M. (1964). Function minimization by conjugate gradients. *Comput. J.* **7** 149–154. MR0187375 https://doi.org/10.1093/comjnl/7.2.149

FOSDICK, B. K. and RAFTERY, A. E. (2014). Regional probabilistic fertility forecasting by modeling between-country correlations. *Demogr. Res.* **30** 1011.

FRENI-STERRANTINO, A., VENTRUCCI, M. and RUE, H. (2018). A note on intrinsic conditional autoregressive models for disconnected graphs. *Spat. Spatio-Tempor. Epidemiol.* **26** 25–34.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GAŁECKI, A. and BURZYKOWSKI, T. (2013). *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. *Springer Texts in Statistics*. Springer, New York. MR3024843 https://doi.org/10.1007/978-1-4614-3900-4

GALLOWAY, M. (2018). CVglasso: Lasso Penalized Precision Matrix Estimation. R package version 1.0.

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

GOLDFARB, D. and IDNANI, A. (1983a). A numerically stable dual method for solving strictly convex quadratic programs. *Math. Program.* **27** 1–33. MR0712108 https://doi.org/10.1007/BF02591962

GOLDFARB, D. and IDNANI, A. (2006). Dual and primal-dual methods for solving strictly convex quadratic programs. In *Numerical Analysis: Proceedings of the Third IIMAS Workshop Held at Cocoyoc, Mexico, January* 1981, pp. 226–239. Springer, Berlin.

HANNAN, E. J. (1973). The asymptotic theory of linear time-series models. *J. Appl. Probab.* **10** 130–145, corrections, ibid. 10 (1973), 913. MR0365960 https://doi.org/10.1017/s0021900200042145

HARSHMAN, R. A. and LUNDY, M. E. (1994). Parafac: Parallel factor analysis. *Comput. Statist. Data Anal.* **18** 39–72.

JOSSE, J. and HUSSON, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *J. Stat. Softw.* **70** 1–31.

KAROLYI, G. A. (1992). Predicting risk: Some new generalizations. *Management Science* **38** 57–74.

KYUNG, M. and GHOSH, S. K. (2010). Maximum likelihood estimation for directional conditionally autoregressive models. *J. Statist. Plann. Inference* **140** 3160–3179. MR2659845 https://doi.org/10.1016/j.jspi.2010.04.012

LEDOIT, O. and WOLF, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance* **10** 603–621.

LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339 https://doi.org/10.1016/S0047-259X(03)00096-4
LEDOIT, O. and WOLF, M. (2022). The power of (non-) linear shrinking: A review and guide to covariance matrix estimation. *J. Financ. Econom.* **20** 187–218.
LEWANDOWSKI, D., KUROWICKA, D. and JOE, H. (2009). Generating random correlation matrices based on Vines and extended onion method. *J. Multivariate Anal.* **100** 1989–2001. MR2543081 https://doi.org/10.1016/j.jmva.2009.04.008
LIECHTY, J. C., LIECHTY, M. W. and MÜLLER, P. (2004). Bayesian correlation estimation. *Biometrika* **91** 1–14. MR2050456 https://doi.org/10.1093/biomet/91.1.1
LIU, H., WANG, L. and ZHAO, T. (2014). Sparse covariance matrix estimation with eigenvalue constraints. *J. Comput. Graph. Statist.* **23** 439–459. MR3215819 https://doi.org/10.1080/10618600.2013.782818
LONGFORD, N. T. and MUTHÉN, B. O. (1992). Factor analysis for clustered observations. *Psychometrika* **57** 581–597. MR1246752 https://doi.org/10.1007/BF02294421
LOPES, H. F., GAMERMAN, D. and SALAZAR, E. (2011). Generalized spatial dynamic factor models. *Comput. Statist. Data Anal.* **55** 1319–1330. MR2741417 https://doi.org/10.1016/j.csda.2010.09.020
LOPES, H. F., SALAZAR, E. and GAMERMAN, D. (2008). Spatial dynamic factor analysis. *Bayesian Anal.* **3** 759–792. MR2469799 https://doi.org/10.1214/08-BA329
LOPUHAÄ, H. P., GARES, V. and RUIZ-GAZEN, A. (2023). S-estimation in linear models with structured covariance matrices. *Ann. Statist.* **51** 2415–2439. MR4682703 https://doi.org/10.1214/23-aos2334
LYONS, R. (1988). Strong laws of large numbers for weakly correlated random variables. *Michigan Math. J.* **35** 353–359. MR0978305 https://doi.org/10.1307/mmj/1029003816
MACNAB, Y. C. (2011). On Gaussian Markov random fields and Bayesian disease mapping. *Stat. Methods Med. Res.* **20** 49–68. MR2767372 https://doi.org/10.1177/0962280210371561
MARTINI, J. W., CROSSA, J., TOLEDO, F. H. and CUEVAS, J. (2020). On Hadamard and Kronecker products in covariance structures for genotype× environment interaction. *The Plant Genome* **13** e20033.
MAYER, T. and ZIGNAGO, S. (2006). Notes on CEPII's distances measures. Electronic resource. https://mpra.ub.uni-muenchen.de/26469/1/MPRA_paper_26469.pdf.
METODIEV, M., PERROT-DOCKÈS, M., OUADAH, S., FOSDICK, B. K, ROBIN, S., LATOUCHE, P. and RAFTERY, A. E (2026). Supplement to "A structured estimator for large covariance matrices in the presence of pairwise and spatial covariates." https://doi.org/10.1214/26-AOAS2183SUPPA, https://doi.org/10.1214/26-AOAS2183SUPPB
METODIEV, M., PERROT-DOCKÈS, M. and ROBIN, S. (2025). scov: Structured Covariances Estimators for Pairwise and Spatial Covariates. R package version 2.0.0.
UNITED NATIONS (2010). *World Population Prospects: the 2010 Revision, Volume I: Comprehensive Tables*. New York, N.Y.
UNITED NATIONS (2024). *World Population Prospects: the 2024 Revision, Volume I: Comprehensive Tables*. New York, N.Y.
POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. MR1723786 https://doi.org/10.1093/biomet/86.3.677
POURAHMADI, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statist. Sci.* **26** 369–387. MR2917961 https://doi.org/10.1214/11-STS358
POURAHMADI, M. (2013). *High-Dimensional Covariance Estimation. Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3235948 https://doi.org/10.1002/9781118573617
PRESTON, S. H., HEUVELINE, P. and GUILLOT, M. (2001). *Measuring and Modeling Population Processes*. Blackwell Publishers, Oxford, U.K.
QIAN, L. (2009). Bayesian semiparametric correlation models for longitudinal data with applications to an HIV/AIDS biomarker study Phd thesis, Univ. California. Available at https://www.proquest.com/docview/304854584?pq-origsite=gscholar&fromope%nview=true&sourcetype=Dissertations%20&%20Theses.
RAFTERY, A. E. (1995). Bayesian model selection in social research. *Sociol. Method.* 111–163.
RAFTERY, A., HOETING, J., VOLINSKY, C., PAINTER, I. and YEUNG, K. Y. (2013a). BMA: Bayesian model averaging. R package version 3.16.1.
RAFTERY, A. E., CHUNN, J. L., GERLAND, P. and ŠEVČÍKOVÁ, H. (2013b). Bayesian probabilistic projections of life expectancy for all countries. *Demography* **50** 777–801.
RAFTERY, A. E., LI, N., ŠEVČÍKOVÁ, H., GERLAND, P. and HEILIG, G. K. (2012). Bayesian probabilistic population projections for all countries. *Proc. Natl. Acad. Sci. USA* **109** 13915–13921.
RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581
SEAMAN, S., GALATI, J., JACKSON, D. and CARLIN, J. (2013). What is meant by "missing at random"? *Statist. Sci.* **28** 257–268. MR3112409 https://doi.org/10.1214/13-sts415

SHANNO, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Math. Comp.* **24** 647–656. MR0274029 https://doi.org/10.2307/2004840

SUN, Y., BABU, P. and PALOMAR, D. P. (2016). Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions. *IEEE Trans. Signal Process.* **64** 3576–3590. MR3515702 https://doi.org/10.1109/TSP.2016.2546222

TASTU, J., PINSON, P. and MADSEN, H. (2013). Space-time scenarios of wind power generation produced using a Gaussian copula with parametrized precision matrix.

THORSON, J. T., SCHEUERELL, M. D., SHELTON, A. O., SEE, K. E., SKAUG, H. J. and KRISTENSEN, K. (2015). Spatial factor analysis: A new tool for estimating joint species distributions and correlations in species range. *Methods Ecol. Evol.* **6** 627–637.

TOKUDA, T., GOODRICH, B., VAN MECHELEN, I., GELMAN, A. and TUERLINCKX, F. (2011). Visualizing distributions of covariance matrices. Columbia Univ, New York. Tech. Rep, 18–18.

TURLACH, B. A. and WEINGESSEL, A. (2019). quadprog: Functions to solve quadratic programming problems. R package version 1.5-8.

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

VER HOEF, J. M., HANKS, E. M. and HOOTEN, M. B. (2018). On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models. *Spat. Stat.* **25** 68–85. MR3809256 https://doi.org/10.1016/j.spasta.2018.04.006

WALL, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *J. Statist. Plann. Inference* **121** 311–324. MR2038824 https://doi.org/10.1016/S0378-3758(03)00111-3

WANG, F. and WALL, M. M. (2003). Generalized common spatial factor model. *Biostatistics* **4** 569–582.

WEI, T. and SIMKO, V. (2021). R package 'corrplot': Visualization of a correlation matrix. (Version 0.92).

WEST, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics* **7** 733–742.

ZHOU, R. and PALOMAR, D. P. (2024). covFactorModel: Covariance matrix estimation via factor models. R package version 0.1.0.

The Annals of Applied Statistics
2026, Vol. 20, No. 2, 1766–1787
https://doi.org/10.1214/26-AOAS2193

THE BRADLEY–TERRY STOCHASTIC BLOCK MODEL

BY LAPO SANTI[1,a] AND NIAL FRIEL[2,b]

[1]*School of Mathematics and Statistics, University College Dublin,* [a]*lapo.santi@ucdconnect.ie*
[2]*Insight Centre for Data Analytics, University College Dublin,* [b]*nial.friel@ucd.ie*

The Bradley–Terry model is widely used for the analysis of pairwise comparison data and, in essence, produces a ranking of the items under comparison. We embed the Bradley–Terry model within a stochastic block model, allowing items to cluster. The resulting Bradley–Terry SBM (BT–SBM) ranks clusters so that items within a cluster share the same tied rank. We develop a fully Bayesian specification in which all quantities—the number of blocks, their strengths, and item assignments—are jointly learned via a fast Gibbs sampler derived through a Thurstonian data augmentation. Despite its efficiency, the sampler yields coherent and interpretable posterior summaries for all model components. Our motivating application analyses men's tennis results from ATP tournaments from the 2000 season up to the 2025 season. We find that the top 105 players can be broadly partitioned into three or four tiers in most seasons. Moreover, the size of the strongest tier was small from the mid-2000s to 2018. Between 2019 and 2022, we observe a transition period characterised by a gradual widening of the top tier, while in more recent seasons (2023–2025) the structure appears to revert to a more elite configuration, coinciding with the rise of dominant players such as Carlos Alcaraz and Jannik Sinner.

REFERENCES

AGRESTI, A. (2013). *Categorical Data Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR3087436

BALDASSARRE, A., DUSSELDORP, E., D'AMBROSIO, A., DE ROOIJ, M. and CONVERSANO, C. (2023). The Bradley-Terry regression trunk approach for modeling preference data with small trees. *Psychometrika* **88** 1443–1465. MR4668575 https://doi.org/10.1007/s11336-022-09882-6

BASINI, F., TSOULI, V., NTZOUFRAS, I. and FRIEL, N. (2023). Assessing competitive balance in the English Premier League for over forty seasons using a stochastic block model. *J. Roy. Statist. Soc., Ser. A* **186** 530–556. MR4719340 https://doi.org/10.1093/jrsssa/qnad007

BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39** 324–345. MR0070925 https://doi.org/10.2307/2334029

CARON, F. and DOUCET, A. (2012). Efficient Bayesian inference for generalized Bradley-Terry models. *J. Comput. Graph. Statist.* **21** 174–196. MR2913362 https://doi.org/10.1080/10618600.2012.638220

CATTELAN, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statist. Sci.* **27** 412–433. MR3012434 https://doi.org/10.1214/12-STS396

CHIANG, W.-L., ZHENG, L., SHENG, Y., ANGELOPOULOS, A. N., LI, T., LI, D., ZHU, B., ZHANG, H., JORDAN, M. I. et al. (2024). Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the* 41*st International Conference on Machine Learning. ICML'*24. JMLR.org. https://doi.org/10.48550/ARXIV.2403.04132

CROON, M. A. and LUIJKX, R. (1993). Latent structure models for ranking data. In *Probability Models and Statistical Analyses for Ranking Data* (*Amherst, MA*, 1990) (M. A. Fligner, J. S. Verducci, J. Berger, S. Fienberg, J. Gani, K. Krickeberg, I. Olkin and B. Singer, eds.) *Lect. Notes Stat.* **80** 53–74. Springer, New York. MR1237201 https://doi.org/10.1007/978-1-4612-2738-0_4

DE BLASI, P., FAVARO, S., LIJOI, A., MENA, R. H., PRÜNSTER, I. and RUGGIERO, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 212–229. https://doi.org/10.1109/TPAMI.2013.217

Key words and phrases. Bradley–Terry model, stochastic block model, ranking data, Bayesian inference, Gibbs sampling, tennis analytics.

GLICKMAN, M. E. (2008). Bayesian locally optimal design of knockout tournaments. *J. Statist. Plann. Inference* **138** 2117–2127. MR2420307 https://doi.org/10.1016/j.jspi.2007.09.007

GOLDSTEIN, H. and SPIEGELHALTER, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *J. Roy. Statist. Soc., Ser. A, Statist. Soc.* **159** 385–443.

GORMLEY, I. C. and MURPHY, T. B. (2009). A grade of membership model for rank data. *Bayesian Anal.* **4** 265–295. MR2507364 https://doi.org/10.1214/09-BA410

GUIVER, J. and SNELSON, E. (2009). *Bayesian Inference for Plackett-Luce Ranking Models* 377–384. https://doi.org/10.1145/1553374.1553423

HATZINGER, R. and MAZANEC, J. A. (2007). Measuring the part worth of the mode of transport in a trip package: An extended Bradley–Terry model for paired-comparison conjoint data. *J. Bus. Res.* **60** 1290–1302. https://doi.org/10.1016/j.jbusres.2007.04.010

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

HSIAO, C. and WANG, I. (2021). Joint clustering and ranking from heterogeneous pairwise comparisons. IEEE international symposium on information theory. In 2021 *IEEE International Symposium on Information Theory* (*ISIT*) 2036–2041. IEEE Press, New York. https://doi.org/10.1109/isit45174.2021.9517936

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. **2** 193–218. https://doi.org/10.1007/bf01908075

JOSEPHS, B. D. and UPTON, J. (2025). Hypergraph Adjusted Plus-Minus: a Network-Based Approach to Player Evaluation in Sports. https://doi.org/10.48550/arXiv.2403.20214

LEGRAMANTI, S., RIGON, T., DURANTE, D. and DUNSON, D. B. (2022). Extended stochastic block models with application to criminal networks. *Ann. Appl. Stat.* **16** 2369–2395. MR4489215 https://doi.org/10.1214/21-aoas1595

LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 715–740. MR2370077 https://doi.org/10.1111/j.1467-9868.2007.00609.x

LOEWEN, P. J., RUBENSON, D. and SPIRLING, A. (2012). Testing the power of arguments in referendums: A Bradley–Terry approach. *Elect. Stud.* **31** 212–221. Special Symposium: Germany's Federal Election September 2009. https://doi.org/10.1016/j.electstud.2011.07.003

LU, C., DURANTE, D. and FRIEL, N. (2026). Zero-inflated stochastic block modelling of efficiency-security trade-offs in weighted criminal networks. *J. Roy. Statist. Soc., Ser. A* **189** 869–897. MR5057953 https://doi.org/10.1093/jrsssa/qnaf029

MASAROTTO, G. and VARIN, C. (2012). The ranking lasso and its application to sport tournaments. *Ann. Appl. Stat.* **6** 1949–1970. MR3058689 https://doi.org/10.1214/12-AOAS581

MEILĂ, M. (2007). Comparing clusterings—an information based distance. *J. Multivariate Anal.* **98** 873–895. MR2325412 https://doi.org/10.1016/j.jmva.2006.11.013

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

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

PEARCE, M. and EROSHEVA, E. A. (2025). Bayesian rank-clustering. *Psychometrika* **90** 904–931. MR4963966 https://doi.org/10.1017/psy.2025.10014

PIANCASTELLI, L. S. C. and FRIEL, N. (2025). The clustered Mallows model. *Stat. Comput.* **35** Paper No. 21, 21. MR4845078 https://doi.org/10.1007/s11222-024-10555-w

PIETTE, J., PHAM, L. and ANAND, S. (2011). Evaluating basketball player performance via statistical network modeling. In *MIT Sloan Sports Analytics Conference* 1–11.

PITMAN, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **30** 245–267. IMS, Hayward, CA. MR1481784 https://doi.org/10.1214/lnms/1215453576

SANTI, L. and FRIEL, N. (2026). Supplement to "The Bradley–Terry stochastic block model." https://doi.org/10.1214/26-AOAS2193SUPPA, https://doi.org/10.1214/26-AOAS2193SUPPB

SCHÖTTL, K., KEINER, M., METZ, V. and KAINZ, P. (2025). The financial break even in professional tennis: From which ranking can you afford your life from professional tennis? *Soc. Sci. Humanit. Open*. [Manuscript submitted for publication. Status: accepted]. https://doi.org/10.2139/ssrn.5206641

SEYMOUR, R. G., SIRL, D., PRESTON, S. P., DRYDEN, I. L., ELLIS, M. J. A., PERRAT, B. and GOULDING, J. (2022). The Bayesian spatial Bradley-Terry model: Urban deprivation modelling in Tanzania. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **71** 288–308. MR4396910 https://doi.org/10.1111/rssc.12532

SHEN, L., AMINI, A., JOSEPHS, N. and LIN, L. (2025). Bayesian community detection for networks with covariates. *Bayesian Anal.* **20** 735–762. MR4942744 https://doi.org/10.1214/24-BA1415

SPEARING, H., TAWN, J., IRONS, D. and PAULDEN, T. (2023). Modeling intransitivity in pairwise comparisons with application to baseball data. *J. Comput. Graph. Statist.* **32** 1383–1392. MR4669254 https://doi.org/10.1080/10618600.2023.2177299

TURNER, H. and FIRTH, D. (2012). Bradley–Terry models in R: The BradleyTerry2 package. *J. Stat. Softw.* **48** 1–21.

TUTZ, G. and SCHAUBERGER, G. (2015). Extended ordered paired comparison models with application to football data from German Bundesliga. *AStA Adv. Stat. Anal.* **99** 209–227. MR3328161 https://doi.org/10.1007/s10182-014-0237-1

VACA-RAMÍREZ, F. and PEIXOTO, T. P. (2022). Systematic assessment of the quality of fit of the stochastic block model for empirical networks. *Phys. Rev. E* **105** Paper No. 054311, 24. MR4445614 https://doi.org/10.1103/physreve.105.054311

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

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

WAINER, J. (2023). A Bayesian Bradley-Terry model to compare multiple ML algorithms on multiple data sets. *J. Mach. Learn. Res.* **24** Paper No. [341], 34. MR4690290

WHELAN, J. T. (2017). Prior distributions for the Bradley–Terry model of paired comparisons. https://doi.org/10.48550/arXiv.1712.05311

WU, R., XU, J., SRIKANT, R., MASSOULIÉ, L., LELARGE, M. and HAJEK, B. (2015). Clustering and inference from pairwise comparisons. In *Proceedings of the* 2015 *ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems* 449–450. ACM, New York. https://doi.org/10.1145/2835776.2835787