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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles

continued

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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles—*Continued from front cover*

THE ANNALS OF APPLIED STATISTICS Vol. 16, No. 3, pp. 1253–2053 September 2022

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

IMS OFFICERS

IMS PUBLICATIONS

The Annals of Applied Statistics **[ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 16, Number 3, September 2022. 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
2022, Vol. 16, No. 3, 1253–1267
https://doi.org/10.1214/21-AOAS1479

MODEL-BASED DISTANCE EMBEDDING WITH APPLICATIONS TO CHROMOSOMAL CONFORMATION BIOLOGY

BY YUPING ZHANG[1,a], DISHENG MAO[1,b] AND ZHENGQING OUYANG[2,c]

[1]*Department of Statistics, University of Connecticut,* [a]*yuping.zhang@uconn.edu,* [b]*disheng.mao@uconn.edu*
[2]*Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts,* [c]*ouyang@schoolph.umass.edu*

Recent development of high-throughput biotechnologies, such as Hi-C, have enabled genome-wide measurement of chromosomal conformation. The interaction signals among genomic loci are contaminated with noises. It remains largely unknown how well the underlying chromosomal conformation can be elucidated, based on massive and noisy measurements. We propose a new model-based distance embedding (MDE) framework, to reveal spatial organizations of chromosomes. The proposed framework is a general methodology, which allows us to link accurate probabilistic models, which characterize biological data properties, to efficiently recovering Euclidean distance matrices from noisy observations. The performance of MDE is shown through numerical experiments inspired by regular helix structure and random movement of chromosomes. The practical merits of MDE are also demonstrated by applications to real Hi-C data from both human and mouse cells which are further validated by gold standard benchmarks.

REFERENCES

ALFAKIH, A. Y., KHANDANI, A. and WOLKOWICZ, H. (1999). Solving Euclidean distance matrix completion problems via semidefinite programming. *Comput. Optim. Appl.* **12** 13–30. MR1704098 https://doi.org/10.1023/A:1008655427845

ARUN, K. S., HUANG, T. S. and BLOSTEIN, S. D. (1987). Least-squares fitting of two 3-d point sets. *IEEE Trans. Pattern Anal. Mach. Intell.* **5** 698–700.

DEKKER, J. and MISTELI, T. (2015). Long-range chromatin interactions. *Cold Spring Harb. Perspect. Biol.* **7** a019356.

DYKSTRA, R. L. (1983). An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.* **78** 837–842. MR0727568

ESKELAND, R., LEEB, M., GRIMES, G. R., KRESS, C., BOYLE, S., SPROUL, D., GILBERT, N., FAN, Y., SKOULTCHI, A. I. et al. (2010). Ring1b compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol. Cell* **38** 452–464.

FANG, H. and O'LEARY, D. P. (2012). Euclidean distance matrix completion problems. *Optim. Methods Softw.* **27** 695–717. MR2946053 https://doi.org/10.1080/10556788.2011.643888

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2001). *The Elements of Statistical Learning*, *Vol.* 1, *Springer Series in Statistics*. Springer, New York.

GAFFKE, N. and MATHAR, R. (1989). A cyclic projection algorithm via duality. *Metrika* **36** 29–54. MR0985010 https://doi.org/10.1007/BF02614077

GAO, Y. and SUN, D. (2010). A majorized penalty approach for calibrating rank constrained correlation matrix problems Technical Report, Department of Mathematics, National Univ. Singapore.

GLUNT, W., HAYDEN, T. L., HONG, S. and WELLS, J. (1990). An alternating projection algorithm for computing the nearest Euclidean distance matrix. *SIAM J. Matrix Anal. Appl.* **11** 589–600. MR1066161 https://doi.org/10.1137/0611042

HAN, S.-P. (1988). A successive projection method. *Math. Program.* **40** 1–14. MR0923692 https://doi.org/10.1007/BF01580719

HU, M., DENG, K., QIN, Z., DIXON, J., SELVARAJ, S., FANG, J., REN, B. and LIU, J. S. (2013). Bayesian inference of spatial organizations of chromosomes. *PLoS Comput. Biol.* **9** e1002893.

Key words and phrases. Spatial organization, dimension reduction, 3D structure, euclidean space.

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

PRICE, C. (1993). Fluorescence in situ hybridization. *Blood Rev.* **7** 127–134.

QI, H.-D. (2013). A semismooth Newton method for the nearest Euclidean distance matrix problem. *SIAM J. Matrix Anal. Appl.* **34** 67–93. MR3032992 https://doi.org/10.1137/110849523

QI, H.-D. and YUAN, X. (2014). Computing the nearest Euclidean distance matrix with low embedding dimensions. *Math. Program.* **147** 351–389. MR3258529 https://doi.org/10.1007/s10107-013-0726-0

THE ENCODE PROJECT CONSORTIUM (2012). An integrated encyclopedia of dna elements in the human genome. *Nature* **489** 57.

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.

ZHANG, L., WAHBA, G. and YUAN, M. (2016). Distance shrinkage and Euclidean embedding via regularized kernel estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 849–867. MR3534353 https://doi.org/10.1111/rssb.12138

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** 40.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1268–1291
https://doi.org/10.1214/21-AOAS1534

CONTRASTIVE LATENT VARIABLE MODELING WITH APPLICATION TO CASE-CONTROL SEQUENCING EXPERIMENTS

BY ANDREW JONES[a], F. WILLIAM TOWNES[b], DIDONG LI[c] AND BARBARA E. ENGELHARDT[d]

Department of Computer Science, Princeton University, [a]*aj13@princeton.edu,* [b]*ftownes@princeton.edu,* [c]*didongli@princeton.edu,* [d]*bee@princeton.edu*

High-throughput RNA-sequencing (RNA-seq) technologies are powerful tools for understanding cellular state. Often, it is of interest to quantify and to summarize changes in cell state that occur between experimental or biological conditions. Differential expression is typically assessed using univariate tests to measure genewise shifts in expression. However, these methods largely ignore changes in transcriptional correlation. Furthermore, there is a need to identify the low-dimensional structure of the gene expression shift to identify collections of genes that change between conditions. Here, we propose contrastive latent variable models designed for count data to create a richer portrait of differential expression in sequencing data. These models disentangle the sources of transcriptional variation in different conditions in the context of an explicit model of variation at baseline. Moreover, we develop a model-based hypothesis testing framework that can test for global and gene subset-specific changes in expression. We evaluate our model through extensive simulations and analyses with count-based gene expression data from perturbation and observational sequencing experiments. We find that our methods effectively summarize and quantify complex transcriptional changes in case-control experimental sequencing data.

REFERENCES

ABID, A., ZHANG, M. J., BAGARIA, V. K. and ZOU, J. (2018). Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nat. Commun.* **9** 1–7.

ADAMSON, B., NORMAN, T. M., JOST, M., CHO, M. Y., NUÑEZ, J. K., CHEN, Y., VILLALTA, J. E., GILBERT, L. A., HORLBECK, M. A. et al. (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167** 1867–1882.

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis. Wiley Publications in Statistics*. Wiley, New York; CRC Press, London. MR0091588

AOSHIMA, M. and YATA, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statist. Sinica* **28** 43–62. MR3752251

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

BOILEAU, P., HEJAZI, N. S. and DUDOIT, S. (2020). Exploring high-dimensional biological data with sparse contrastive principal component analysis. *Bioinformatics* **36** 3422–3430. https://doi.org/10.1093/bioinformatics/btaa176

CAI, T., LIU, W. and XIA, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.* **108** 265–277. MR3174618 https://doi.org/10.1080/01621459.2012.758041

CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P. A. and WILLSKY, A. S. (2009). Sparse and low-rank matrix decompositions. *IFAC Proc. Vol.* **42** 1493–1498.

GTEX CONSORTIUM (2017). Genetic effects on gene expression across human tissues. *Nature* **550** 204.

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

Key words and phrases. Latent variable models, RNA sequencing, differential expression, contrastive models, case-control data.

DELMANS, M. and HEMBERG, M. (2016). Discrete distributional differential expression (D3E)–a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinform.* **17** 110. https://doi.org/10.1186/s12859-016-0944-6

DILLON, J. V., LANGMORE, I., TRAN, D., BREVDO, E., VASUDEVAN, S., MOORE, D., PATTON, B., ALEMI, A., HOFFMAN, M. et al. (2017). Tensorflow distributions. Preprint. Available at arXiv:1711.10604.

DING, J., CONDON, A. and SHAH, S. P. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* **9** 1–13.

DIXIT, A., PARNAS, O., LI, B., CHEN, J., FULCO, C. P., JERBY-ARNON, L., MARJANOVIC, N. D., DIONNE, D., BURKS, T. et al. (2016). Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167** 1853–1866.

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

GLASS, K., HUTTENHOWER, C., QUACKENBUSH, J. and YUAN, G.-C. (2013). Passing messages between biological networks to refine predicted interactions. *PLoS ONE* **8** e64832.

GOODMAN, S. N. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med.* **130** 1005–1013.

HAFEMEISTER, C. and SATIJA, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20** 1–15.

HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.* **14** 1303–1347. MR3081926

ISHII, A., YATA, K. and AOSHIMA, M. (2019). Equality tests of high-dimensional covariance matrices under the strongly spiked eigenvalue model. *J. Statist. Plann. Inference* **202** 99–111. MR3926765 https://doi.org/10.1016/j.jspi.2019.02.002

JOHNSTONE, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *Ann. Statist.* **36** 2638–2716. MR2485010 https://doi.org/10.1214/08-AOS605

JONES, A., TOWNES, F. W, LI, D. and ENGELHARDT, B. E (2022). Supplement to "Contrastive latent variable modeling with application to case-control sequencing experiments." https://doi.org/10.1214/21-AOAS1534SUPPA, https://doi.org/10.1214/21-AOAS1534SUPPB

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

KHARCHENKO, P. V., SILBERSTEIN, L. and SCADDEN, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11** 740–742.

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. Preprint. Available at arXiv:1412.6980.

KINKER, G. S., GREENWALD, A. C., TAL, R., ORLOVA, Z., CUOCO, M. S., MCFARLAND, J. M., WARREN, A., RODMAN, C., ROTH, J. A. et al. (2020). Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.* **52** 1208–1218.

KORTHAUER, K. D., CHU, L.-F., NEWTON, M. A., LI, Y., THOMSON, J., STEWART, R. and KENDZIORSKI, C. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* **17** 222.

LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105** 18718–18723.

LI, J. and CHEN, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *Ann. Statist.* **40** 908–940. MR2985938 https://doi.org/10.1214/12-AOS993

LI, D., JONES, A. and ENGELHARDT, B. (2020). Probabilistic contrastive principal component analysis. Preprint. Available at arXiv:2012.07977.

LIBERZON, A., BIRGER, C., THORVALDSDÓTTIR, H., GHANDI, M., MESIROV, J. P. and TAMAYO, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Syst.* **1** 417–425.

LOPEZ, R., REGIER, J., COLE, M. B., JORDAN, M. I. and YOSEF, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15** 1053–1058. https://doi.org/10.1038/s41592-018-0229-2

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.

MCFARLAND, J. M., PAOLELLA, B. R., WARREN, A., GEIGER-SCHULLER, K., SHIBUE, T., ROTHBERG, M., KUKSENKO, O., COLGAN, W. N., JONES, A. et al. (2020). Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat. Commun.* **11** 1–15.

MIAO, Z., DENG, K., WANG, X. and ZHANG, X. (2018). DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* **34** 3223–3224.

NABAVI, S., SCHMOLZE, D., MAITITUOHETI, M., MALLADI, S. and BECK, A. H. (2016). EMDomics: A robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics* **32** 533–541.

O'BRIEN, P. C. (1992). Robust procedures for testing equality of covariance matrices. *Biometrics* 819–827.

QIU, X., HILL, A., PACKER, J., LIN, D., MA, Y.-A. and TRAPNELL, C. (2017). Single-cell mRNA quantification and differential analysis with census. *Nat. Methods* **14** 309–315.

ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.

SEVERSON, K. A., GHOSH, S. and NG, K. (2019). Unsupervised learning with contrastive latent variable models. In *Proceedings of the AAAI Conference on Artificial Intelligence* **33** 4862–4869.

SRIVASTAVA, M. S. and YANAGIHARA, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *J. Multivariate Anal.* **101** 1319–1329. MR2609494 https://doi.org/10.1016/j.jmva.2009.12.010

STUART, J. M., SEGAL, E., KOLLER, D. and KIM, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302** 249–255.

TOWNES, F. W., HICKS, S. C., ARYEE, M. J. and IRIZARRY, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model. *Genome Biol.* **20** 1–16.

VASSILEV, L. T., VU, B. T., GRAVES, B., CARVAJAL, D., PODLASKI, F., FILIPOVIC, Z., KONG, N., KAMMLOTT, U., LUKACS, C. et al. (2004). In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* **303** 844–848.

XIA, Y., CAI, T. and CAI, T. T. (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika* **102** 247–266. MR3371002 https://doi.org/10.1093/biomet/asu074

YOUNG, M. D., MITCHELL, T. J., BRAGA, F. A. V., TRAN, M. G., STEWART, B. J., FERDINAND, J. R., COLLORD, G., BOTTING, R. A., POPESCU, D.-M. et al. (2018). Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* **361** 594–599.

ZAPPIA, L., PHIPSON, B. and OSHLACK, A. (2017). Splatter: Simulation of single-cell RNA sequencing data. *Genome Biol.* **18** 1–15.

WANG, and LI, and NELSON, E. and NABAVI, (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinform.* **20** 1–16. https://doi.org/10.1186/s12859-019-2599-6

ZHU, L., LEI, J., DEVLIN, B. and ROEDER, K. (2017). Testing high-dimensional covariance matrices, with application to detecting schizophrenia risk genes. *Ann. Appl. Stat.* **11** 1810–1831. MR3709579 https://doi.org/10.1214/17-AOAS1062

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527 https://doi.org/10.1198/106186006X113430

ZOU, J. Y., HSU, D. J., PARKES, D. C. and ADAMS, R. P. (2013). Contrastive learning using spectral methods. *Adv. Neural Inf. Process. Syst.* **26** 2238–2246.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1292–1312
https://doi.org/10.1214/21-AOAS1537

B-SCALING: A NOVEL NONPARAMETRIC DATA FUSION METHOD

BY YIWEN LIU [1,a], XIAOXIAO SUN [1,b], WENXUAN ZHONG[2,c] AND BING LI[3,d]

[1]*Department of Epidemiology and Biostatistics, University of Arizona,* [a]*yiwenliu@arizona.edu,* [b]*xiaosun@arizona.edu*
[2]*Department of Statistics, University of Georgia,* [c]*wenxuan@uga.edu*
[3]*Department of Statistics, Pennsylvania State University,* [d]*bing@stat.psu.edu*

Very often for the same scientific question, there may exist different techniques or experiments that measure the same numerical quantity. Historically, various methods have been developed to exploit the information within each type of data independently. However, statistical data fusion methods that could effectively integrate multisource data under a unified framework are lacking. In this paper we propose a novel data fusion method, called B-scaling, for integrating multisource data. Consider K measurements that are generated from different sources but measure the same latent variable through some linear or nonlinear ways. We seek to find a representation of the latent variable, named B-mean, which captures the common information contained in the K measurements while taking into account the nonlinear mappings between them and the latent variable. We also establish the asymptotic property of the B-mean and apply the proposed method to integrate multiple histone modifications and DNA methylation levels for characterizing epigenomic landscape. Both numerical and empirical studies show that B-scaling is a powerful data fusion method with broad applications.

REFERENCES

AMIN, V., HARRIS, R. A., ONUCHIC, V., JACKSON, A. R., CHARNECKI, T., PAITHANKAR, S., SUBRAMANIAN, S. L., RIEHLE, K., COARFA, C. et al. (2015). Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. *Nat. Commun.* **6** 1–10.

BERNSTEIN, B. E., STAMATOYANNOPOULOS, J. A., COSTELLO, J. F., REN, B., MILOSAVLJEVIC, A., MEISSNER, A., KELLIS, M., MARRA, M. A., BEAUDET, A. L. et al. (2010). The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28** 1045–1048.

BIRD, A. (2007). Perceptions of epigenetics. *Nature* **447** 396–398. https://doi.org/10.1038/nature05913

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*: *A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press/CRC, Boca Raton, FL. MR2243417 https://doi.org/10.1201/9781420010138

COX, T. F. and COX, M. A. A. (2000). *Multidimensional Scaling*. Chapman and Hall/CRC.

EGGER, G., LIANG, G., APARICIO, A. and JONES, P. A. (2004). Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **429** 457.

ERNST, J. and KELLIS, M. (2012). Chromhmm: Automating chromatin-state discovery and characterization. *Nat. Methods* **9** 215–216.

ESTELLER, M. (2008). Epigenetics in cancer. *N. Engl. J. Med.* **358** 1148–1159.

FEINBERG, A. P., KOLDOBSKIY, M. A. and GÖNDÖR, A. (2016). Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat. Rev. Genet.* **17** 284–299.

FERNHOLZ, L. T. (2012). *Von Mises Calculus for Statistical Functionals*. *Lecture Notes in Statistics* **19**. Springer Science & Business Media, New York.

FULLER, W. A. (2009). *Measurement Error Models*. *Wiley Series in Probability and Statistics*, *Vol.* 305. Wiley, Hoboken, NJ.

GOMEZ-CABRERO, D., ABUGESSAISA, I., MAIER, D., TESCHENDORFF, A., MERKENSCHLAGER, M., GISEL, A., BALLESTAR, E., BONGCAM-RUDLOFF, E., CONESA, A. et al. (2014). Data integration in the era of omics: Current and future challenges. *BMC Syst. Biol.* **8** I1.

HALL, D. L. and LLINAS, J. (1997). An introduction to multisensor data fusion. *Proc. IEEE* **85** 6–23.

HALL, D. L. and MCMULLEN, S. A. (2004). *Mathematical Techniques in Multisensor Data Fusion*. Artech House.

Key words and phrases. Data fusion, multisource data, generalized eigenvalue problem, epigenetics.

HAMID, J. S., HU, P., ROSLIN, N. M., LING, V., GREENWOOD, C. M. and BEYENE, J. (2009). Data integration in genetics and genomics: Methods and challenges. In *Human Genomics and Proteomics*: *HGP*.

HE, X., SHEN, L. and SHEN, Z. (2001). A data-adaptive knot selection scheme for fitting splines. *IEEE Signal Process. Lett.* **8** 137–139.

KHALEGHI, B., KHAMIS, A., KARRAY, F. O. and RAZAVI, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Inf. Fusion* **14** 28–44.

KOLLO, T. and VON ROSEN, D. (2005). *Advanced Multivariate Statistics with Matrices*. *Mathematics and Its Applications* (*New York*) **579**. Springer, Dordrecht. MR2162145 https://doi.org/10.1007/1-4020-3419-9

KRUSKAL, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29** 1–27. MR0169712 https://doi.org/10.1007/BF02289565

KUNDAJE, A., MEULEMAN, W., ERNST, J., BILENKY, M., BERNSTEIN, B. E., COSTELLO, J. F., ECKER, J. R., HIRST, M., MEISSNER, A. et al. (2015). February. Integrative analysis of 111 reference human epigenomes. *Nature* **518** 317–330.

LEE, L., WANG, K., LI, G., XIE, Z., WANG, Y., XU, J., SUN, S., POCALYKO, D., BHAK, J. et al. (2011). Liverome: A curated database of liver cancer-related gene signatures with self-contained context information. *BMC Genomics* **12** S3.

LETUNIC, I., COPLEY, R. R., SCHMIDT, S., CICCARELLI, F. D., DOERKS, T., SCHULTZ, J., PONTING, C. P. and BORK, P. (2004). Smart 4.0: Towards genomic data integration. *Nucleic Acids Res.* **32** D142–D144.

LI, B. (2018). Linear operator-based statistical analysis: A useful paradigm for big data. *Canad. J. Statist.* **46** 79–103. MR3767167 https://doi.org/10.1002/cjs.11329

LI, B. and WANG, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102** 997–1008. MR2354409 https://doi.org/10.1198/016214507000000536

LIU, Y. SUN, X. ZHONG, W. and LI, B. (2022). Supplement to "B-scaling: A novel nonparametric data fusion method." https://doi.org/10.1214/21-AOAS1537SUPPA, https://doi.org/10.1214/21-AOAS1537SUPPB

MENG, C., ZELEZNIK, O. A., THALLINGER, G. G., KUSTER, B., GHOLAMI, A. M. and CULHANE, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* **17** 628–641.

PORTELA, A. and ESTELLER, M. (2010). Epigenetic modifications and human disease. *Nat. Biotechnol.* **28** 1057–1068.

RAY, B., LIU, W. and FENYÖ, D. (2017). Adaptive multiview nonnegative matrix factorization algorithm for integration of multimodal biomedical data. *Cancer Inform.* **16**.

REUTER, J. A., SPACEK, D. V. and SNYDER, M. P. (2015). High-throughput sequencing technologies. *Mol. Cell* **58** 586–597.

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.

ROMANOSKI, C. E., GLASS, C. K., STUNNENBERG, H. G., WILSON, L. and ALMOUZNI, G. (2015). February. Epigenomics: Roadmap for regulation. *Nature* **518** 314–316.

SHEAFFER, K. L., KIM, R., AOKI, R., ELLIOTT, E. N., SCHUG, J., BURGER, L., SCHÜBELER, D. and KAESTNER, K. H. (2014). DNA methylation is required for the control of stem cell differentiation in the small intestine. *Genes Dev.* **28** 652–664.

THE ENCODE PROJECT CONSORTIUM (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57.

TORGERSON, W. S. (1952). Multidimensional scaling. I. Theory and method. *Psychometrika* **17** 401–419. MR0054219 https://doi.org/10.1007/BF02288916

WALTZ, E., LLINAS, J. et al. (1990). *Multisensor Data Fusion* **685**. Artech House, Boston, MA.

YUAN, Y., CHEN, N. and ZHOU, S. (2013). Adaptive B-spline knot selection using multi-resolution basis set. *IIE Trans.* **45** 1263–1277.

YUAN, G.-C., MA, P., ZHONG, W. and LIU, J. S. (2006). Statistical assessment of the global regulatory role of histone acetylation in saccharomyces cerevisiae. *Genome Biol.* **7** R70.

ZANG, C., WANG, T., DENG, K., LI, B., HU, S., QIN, Q., XIAO, T., ZHANG, S., MEYER, C. A. et al. (2016). High-dimensional genomic data bias correction and data integration using MANCIE. *Nat. Commun.* **7** 1–8.

ZHANG, S., LI, Q., LIU, J. and ZHOU, X. J. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics* **27** i401–i409.

ZHANG, S., LIU, C.-C., LI, W., SHEN, H., LAIRD, P. W. and ZHOU, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* **40** 9379–9391.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1313–1332
https://doi.org/10.1214/21-AOAS1540

GRAPH LINK PREDICTION IN COMPUTER NETWORKS USING POISSON MATRIX FACTORISATION

BY FRANCESCO SANNA PASSINO[1,a], MELISSA J. M. TURCOTTE[2,c] AND NICHOLAS A. HEARD[1,b]

[1]*Department of Mathematics, Imperial College London,* [a]*f.sannapassino@imperial.ac.uk,* [b]*n.heard@imperial.ac.uk*
[2]*Microsoft 365 Defender, Microsoft Corporation,* [c]*melissa.turcotte@microsoft.com*

Graph link prediction is an important task in cybersecurity: relationships between entities within a computer network, such as users interacting with computers or system libraries and the corresponding processes that use them, can provide key insights into adversary behaviour. Poisson matrix factorisation (PMF) is a popular model for link prediction in large networks, particularly useful for its scalability. In this article PMF is extended to include scenarios that are commonly encountered in cybersecurity applications. Specifically, an extension is proposed to explicitly handle binary adjacency matrices and include known categorical covariates associated with the graph nodes. A seasonal PMF model is also presented to handle seasonal networks. To allow the methods to scale to large graphs, variational methods are discussed for performing fast inference. The results show an improved performance over the standard PMF model and other statistical network models.

REFERENCES

ACHARYA, A., TEFFER, D., HENDERSON, J., TYLER, M., ZHOU, M. and GHOSH, J. (2015). Gamma process Poisson factorization for joint modeling of network and documents. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases* **1** 283–299.

ADOMAVICIUS, G. and TUZHILIN, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **17** 734–749.

AGARWAL, D., ZHANG, L. and MAZUMDER, R. (2011). Modeling item-item similarities for personalized recommendations on Yahoo! Front page. *Ann. Appl. Stat.* **5** 1839–1875. MR2884924 https://doi.org/10.1214/11-AOAS475

AMIT, I., MATHERLY, J., HEWLETT, W., XU, Z., MESHI, Y. and WEINBERGER, Y. (2019). Machine learning in cyber-security—problems, challenges and data sets. In *AAAI-19 Workshop on Engineering Dependable and Secure Machine Learning Systems*.

ANDERSON, B., VEJMAN, M., MCGREW, D. and PAUL, S. (2018). Towards Generalisable Network Threat Detection. In *Data Science for Cyber-Security* 77–94 4. World Scientific, Singapore.

ATHREYA, A., FISHKIND, D. E., TANG, M., PRIEBE, C. E., PARK, Y., VOGELSTEIN, J. T., LEVIN, K., LYZINSKI, V., QIN, Y. et al. (2017). Statistical inference on random dot product graphs: A survey. *J. Mach. Learn. Res.* **18** Paper No. 226, 92. MR3827114

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

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

CANNY, J. (2004). GaP: A factor model for discrete data. In *Proceedings of the 27th Annual International ACM SIGIR Conference*, *SIGIR'*04 122–129.

CEMGIL, A. T. (2009). Bayesian inference for nonnegative matrix factorisation models. *Comput. Intell. Neurosci.* 785152. https://doi.org/10.1155/2009/785152

CHANEY, A. J. B., BLEI, D. M. and ELIASSI-RAD, T. (2015). A probabilistic model for using social networks in personalized item recommendation. In *Proceedings of the 9th ACM Conference on Recommender Systems*, *RecSys'*15 43–50. ACM, New York.

CHARLIN, L., RANGANATH, R., MCINERNEY, J. and BLEI, D. M. (2015). Dynamic Poisson factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems* 155–162. ACM, New York.

Key words and phrases. Anomaly detection, dynamic networks, new link prediction, Poisson matrix factorisation, statistical cybersecurity, variational inference.

CHEN, B., LI, F., CHEN, S., HU, R. and CHEN, L. (2017). Link prediction based on non-negative matrix factorization. *PLoS ONE* **12** 1–18.

CLAUSET, A., MOORE, C. and NEWMAN, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature* **453**.

DA SILVA, E. D. S., LANGSETH, H. and RAMAMPIARO, H. (2017). Content-based social recommendation with Poisson matrix factorization. In *Machine Learning and Knowledge Discovery in Databases* 530–546.

DAI, B., WANG, J., SHEN, X. and QU, A. (2019). Smooth neighborhood recommender systems. *J. Mach. Learn. Res.* **20** Paper No. 16, 24. MR3911423

DHILLON, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, *KDD'*01 269–274. ACM, New York, NY, USA.

DUNLAVY, D. M., KOLDA, T. G. and ACAR, E. (2011). Temporal link prediction using matrix and tensor factorizations. *ACM Trans. Knowl. Discov. Data* **5**.

DUNSON, D. B. and HERRING, A. H. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics* **6** 11–25.

FIENBERG, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *J. Comput. Graph. Statist.* **21** 825–839. MR3005799 https://doi.org/10.1080/10618600.2012.738106

FITHIAN, W. and MAZUMDER, R. (2018). Flexible low-rank statistical modeling with missing data and side information. *Statist. Sci.* **33** 238–260. MR3797712 https://doi.org/10.1214/18-STS642

GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.

GOPALAN, P., CHARLIN, L. and BLEI, D. M. (2014). Content-based recommendations with Poisson factorization. In *Proceedings of the* 27*th International Conference on Neural Information Processing Systems*, *NIPS'*14 **2** 3176–3184. MIT Press, Cambridge.

GOPALAN, P., HOFMAN, J. M. and BLEI, D. M. (2015). Scalable recommendation with hierarchical Poisson factorization. In *Proceedings of the* 31*st Conference on Uncertainty in Artificial Intelligence*, *UAI'*15 326–335. AUAI Press, Arlington, VA, USA.

HEARD, N. A., RUBIN-DELANCHY, P. T. G. and LAWSON, D. J. (2014). Filtering automated polling traffic in computer network flow data. In *Proceedings—*2014 *IEEE Joint Intelligence and Security Informatics Conference*, *JISIC* 2014 268–271.

HEARD, N. A., ADAMS, N., RUBIN-DELANCHY, P. and TURCOTTE, M. (2018). *Data Science for Cyber-Security*. World Scientific, (Europe).

HERNÁNDEZ-LOBATO, J. M., HOULSBY, N. and GHAHRAMANI, Z. (2014). Stochastic inference for scalable probabilistic modeling of binary matrices. In *Proceedings of the* 31*st International Conference on Machine Learning*, *ICML'*14 **32** II–379–II–387.

HOFF, P. D. (2005). Bilinear mixed-effects models for dyadic data. *J. Amer. Statist. Assoc.* **100** 286–295. MR2156838 https://doi.org/10.1198/016214504000001015

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262 https://doi.org/10.1198/016214502388618906

HOSSEINI, S., KHODADADI, A., ALIZADEH, K., ARABZADEH, A., FARAJTABAR, M., ZHA, H. and RABIEE, H. R. R. (2018). Recurrent Poisson factorization for temporal recommendation. *IEEE Trans. Knowl. Data Eng.*.

HUGGINS, J. H., CAMPBELL, T., KASPRZAK, M. and BRODERICK, T. (2019). Scalable Gaussian process inference with finite-data mean and variance guarantees. In *Proceedings of Machine Learning Research* **89** 796–805.

JESKE, D. R., STEVENS, N. T., TARTAKOVSKY, A. G. and WILSON, J. D. (2018). Statistical methods for network surveillance. *Appl. Stoch. Models Bus. Ind.* **34** 425–445. MR3845890 https://doi.org/10.1002/asmb.2326

JOHNSON, C. C. (2014). Logistic matrix factorization for implicit feedback data. In *Proceedings of the NIPS* 2014 *Workshop on Distributed Machine Learning and Matrix Computations*.

KHANNA, R., ZHANG, L., AGARWAL, D. and CHEN, B. C. (2013). Parallel matrix factorization for binary response. In *IEEE International Conference on Big Data* 2013 430–438.

KIM, B., LEE, K. H., XUE, L. and NIU, X. (2018). A review of dynamic network models with latent variables. *Stat. Surv.* **12** 105–135. MR3850294 https://doi.org/10.1214/18-SS121

KIVELÄ, M., ARENAS, A., BARTHELEMY, M., GLEESON, J. P., MORENO, Y. and PORTER, M. A. (2014). Multilayer networks. *J. Complex Netw.* **2** 203–271.

KUMAR, R. S. S., WICKER, A. and SWANN, M. (2017). Practical machine learning for cloud intrusion detection: Challenges and the way forward. In *Proceedings of the* 10*th ACM Workshop on Artificial Intelligence and Security*, *AISec'*17 81–90.

LIBEN-NOWELL, D. and KLEINBERG, J. (2007). The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.* **58** 1019–1031.

LÜ, L. and ZHOU, T. (2011). Link prediction in complex networks: A survey. *Phys. A, Stat. Mech. Appl.* **390** 1150–1170.

MENON, A. K. and ELKAN, C. (2011). Link prediction via matrix factorization. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD* 2011, *Part II* 437–452. Springer Berlin Heidelberg, Berlin, Heidelberg.

METELLI, S. and HEARD, N. (2019). On Bayesian new edge prediction and anomaly detection in computer networks. *Ann. Appl. Stat.* **13** 2586–2610. MR4037442 https://doi.org/10.1214/19-aoas1286

NAKAJIMA, S., SUGIYAMA, M. and TOMIOKA, R. (2010). Global analytic solution for variational Bayesian matrix factorization. In *Advances in Neural Information Processing Systems* 23 1768–1776.

NEIL, J., HASH, C., BRUGH, A., FISK, M. and STORLIE, C. B. (2013). Scan statistics for the online detection of locally anomalous subgraphs. *Technometrics* **55** 403–414. MR3176546 https://doi.org/10.1080/00401706.2013.822830

NGUYEN, J. and ZHU, M. (2013). Content-boosted matrix factorization techniques for recommender systems. *Stat. Anal. Data Min.* **6** 286–301. MR3092059 https://doi.org/10.1002/sam.11184

PAPASTAMOULIS, P. and NTZOUFRAS, I. (2020). On the identifiability of Bayesian factor analytic models. Available at arXiv:2004.05105.

PAQUET, U. and KOENIGSTEIN, N. (2013). One-class collaborative filtering with random graphs. In *Proceedings of the* 22*nd International Conference on World Wide Web*, *WWW'*13 999–1008. ACM, New York, NY, USA.

SALAKHUTDINOV, R. and MNIH, A. (2007). Probabilistic matrix factorization. In *Proceedings of the* 20*th International Conference on Neural Information Processing Systems*, *NIPS'*07 1257–1264.

SALTER-TOWNSHEND, M. and MURPHY, T. B. (2013). Variational Bayesian inference for the latent position cluster model for network data. *Comput. Statist. Data Anal.* **57** 661–671. MR2981116 https://doi.org/10.1016/j.csda.2012.08.004

SANNA PASSINO, F., TURCOTTE, M. J. and HEARD, N. A. (2022). Supplement to "Graph link prediction in computer networks using Poisson matrix factorisation." https://doi.org/10.1214/21-AOAS1540SUPPA, https://doi.org/10.1214/21-AOAS1540SUPPB

SCHEIN, A., PAISLEY, J., BLEI, D. M. and WALLACH, H. (2015). Bayesian Poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1045–1054. ACM, New York.

SCHEIN, A., ZHOU, M., BLEI, D. M. and WALLACH, H. (2016). Bayesian Poisson Tucker decomposition for learning the structure of international relations. In *Proceedings of the* 33*rd International Conference on Machine Learning*, *New York*, *NY*, *USA*.

SEEGER, M. and BOUCHARD, G. (2012). Fast variational Bayesian inference for non-conjugate matrix factorization models. In *Artificial Intelligence and Statistics* 1012–1018.

SEWELL, D. K. and CHEN, Y. (2015). Latent space models for dynamic networks. *J. Amer. Statist. Assoc.* **110** 1646–1657. MR3449061 https://doi.org/10.1080/01621459.2014.988214

SINGH, A. P. and GORDON, G. J. (2008). Relational learning via collective matrix factorization. In *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *KDD'*08 650–658. ACM, New York, NY, USA.

TURCOTTE, M. J. M., KENT, A. D. and HASH, C. (2018). Unified Host and Network Data Set. In *Data Science for Cyber-Security* 1–22 1. World Scientific, Singapore.

TURCOTTE, M., MOORE, J., HEARD, N. A. and MCPHALL, A. (2016). Poisson factorization for peer-based anomaly detection. In 2016 *IEEE Conference on Intelligence and Security Informatics* (*ISI*) 208–210. https://doi.org/10.1109/ISI.2016.7745472

WU, Z., PAN, S., CHEN, F., LONG, G., ZHANG, C. and YU, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **32** 4–24. MR4205495

ZHANG, M. and CHEN, Y. (2018). Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems* 31 5165–5175.

ZHANG, W. and WANG, J. (2015). A collective Bayesian Poisson factorization model for cold-start local event recommendation. In *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1455–1464.

ZHOU, M. (2015). Infinite edge partition models for overlapping community detection and link prediction. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, *AISTATS*.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1333–1358
https://doi.org/10.1214/21-AOAS1541

MATRIX COMPLETION METHODS FOR THE TOTAL ELECTRON CONTENT VIDEO RECONSTRUCTION

BY HU SUN[1,a], ZHIJUN HUA[1,b], JIAEN REN[2,e], SHASHA ZOU[2,f], YUEKAI SUN[1,c] AND YANG CHEN[1,d]

[1]*Department of Statistics, University of Michigan, Ann Arbor,* [a]*husun@umich.edu,* [b]*zhijunh@umich.edu,* [c]*yuekai@umich.edu,* [d]*ychenang@umich.edu*

[2]*Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor,* [e]*jiaenren@umich.edu,* [f]*shashaz@umich.edu*

The total electron content (TEC) maps can be used to estimate the signal delay of GPS due to the ionospheric electron content between a receiver and satellite. This delay can result in GPS positioning error. Thus, it is important to monitor the TEC maps. The observed TEC maps have big patches of missingness in the ocean and scattered small areas of missingness on the land. In this paper we propose several extensions of existing matrix completion algorithms to achieve TEC map reconstruction, accounting for spatial smoothness and temporal consistency while preserving important structures of the TEC maps. We call the proposed method video imputation with softImpute, temporal smoothing and auxiliary data (VISTA). Numerical simulations that mimic patterns of real data are given. We show that our proposed method achieves better reconstructed TEC maps, as compared to existing methods in literature. Our proposed computational algorithm is general and can be readily applied for other problems besides TEC map reconstruction.

REFERENCES

AA, E., HUANG, W., LIU, S., RIDLEY, A., ZOU, S., SHI, L., CHEN, Y., SHEN, H., YUAN, T. et al. (2018). Midlatitude plasma bubbles over China and adjacent areas during a magnetic storm on 8 September 2017. *Space Weather* **16** 321–331.

AA, E., ZOU, S., RIDLEY, A., ZHANG, S., COSTER, A. J., ERICKSON, P. J., LIU, S. and REN, J. (2019). Merging of storm time midlatitude traveling ionospheric disturbances and equatorial plasma bubbles. *Space Weather* **17** 285–298.

ABDU, M. A. (2019). Day-to-day and short-term variabilities in the equatorial plasma bubble/spread F irregularity seeding and development. *Progress in Earth and Planetary Science* **6** 1–22.

ACAR, E., DUNLAVY, D. M., KOLDA, T. G. and MØRUP, M. (2011). Scalable tensor factorizations for incomplete data. *Chemom. Intell. Lab. Syst.* **106** 41–56.

BARNES, C., SHECHTMAN, E., FINKELSTEIN, A. and GOLDMAN, D. B. (2009). PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* **28** 24.

BASU, S., GROVES, K. M., BASU, S. and SULTAN, P. J. (2002). Specification and forecasting of scintillations in communication/navigation links: Current status and future plans. *J. Atmos. Sol.-Terr. Phys.* **64** 1745–1754.

BELL, R. M., KOREN, Y. and VOLINSKY, C. (2007). The bellkor solution to the Netflix prize. *KorBell Team's Report to Netflix*.

BENNETT, J. and LANNING, S. (2007). The Netflix prize. In *Proceedings of the KDD Cup Workshop* 2007 3–6. ACM, New York.

BOARD, S. S. (1997). *Space Weather*: *A Research Perspective*. The National Academies Press, Washington, DC.

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

CANDÈS, E. J. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theory* **56** 2053–2080. MR2723472 https://doi.org/10.1109/TIT.2010.2044061

CHEN, Z. and CICHOCKI, A. (2005). Nonnegative matrix factorization with temporal smoothness and/or spatial decorrelation constraints. Laboratory for Advanced Brain Signal Processing, RIKEN, Tech. Rep. 68.

Key words and phrases. Matrix completion, temporal smoothing, auxiliary data, regularization, total electron content (TEC).

CHEN, C., HE, B. and YUAN, X. (2012). Matrix completion via an alternating direction method. *IMA J. Numer. Anal.* **32** 227–245. MR2875250 https://doi.org/10.1093/imanum/drq039

CONKER, R. S., EL-ARINI, M. B., HEGARTY, C. J. and HSIAO, T. (2003). Modeling the effects of ionospheric scintillation on GPS/satellite-based augmentation system availability. *Radio Science* **38** 1–1.

EFRON, B. and MORRIS, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *Ann. Statist.* **4** 22–32. MR0394960

FOSTER, J., COSTER, A., ERICKSON, P., HOLT, J., LIND, F., RIDEOUT, W., MCCREADY, M., VAN EYKEN, A., BARNES, R. et al. (2005). Multiradar observations of the polar tongue of ionization. *J. Geophys. Res.* **110**.

GIMÉNEZ-FEBRER, P., PAGÈS-ZAMORA, A. and GIANNAKIS, G. B. (2019). Matrix completion and extrapolation via kernel regression. *IEEE Trans. Signal Process.* **67** 5004–5017. MR4016859 https://doi.org/10.1109/TSP.2019.2932875

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

HERNÁNDEZ-PAJARES, M., JUAN, J. M., SANZ, J., ORUS, R., GARCIA-RIGO, A., FELTENS, J., KOMJATHY, A., SCHAER, S. C. and KRANKOWSKI, A. (2009). The IGS VTEC maps: A reliable source of ionospheric information since 1998. *Journal of Geodesy* **83** 263–275.

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

HUANG, J.-B., KANG, S. B., AHUJA, N. and KOPF, J. (2014). Image completion using planar structure guidance. *ACM Transactions on Graphics* (*TOG*) **33** 1–10.

JAIN, P., NETRAPALLI, P. and SANGHAVI, S. (2013). Low-rank matrix completion using alternating minimization (extended abstract). In *STOC'*13*—Proceedings of the* 2013 *ACM Symposium on Theory of Computing* 665–674. ACM, New York. MR3210828 https://doi.org/10.1145/2488608.2488693

KIM, H. and PARK, H. (2008a). Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM J. Matrix Anal. Appl.* **30** 713–730. MR2421467 https://doi.org/10.1137/07069239X

KIM, J. and PARK, H. (2008b). Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In 2008 *Eighth IEEE International Conference on Data Mining* 353–362. IEEE, New York.

KOREN, Y. (2008). Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 426–434.

KOREN, Y. (2009). The bellkor solution to the Netflix grand prize. *Netflix Prize Documentation* **81** 1–10.

KOREN, Y., BELL, R. and VOLINSKY, C. (2009). Matrix factorization techniques for recommender systems. *Computer* **42** 30–37.

LEE, D. and SEUNG, H. S. (2000). Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.* **13** 556–562.

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

LI, X. P., HUANG, L., SO, H. C. and ZHAO, B. (2019). A survey on matrix completion: Perspective of signal processing. Preprint. Available at arXiv:1901.10885.

LIU, J., MUSIALSKI, P., WONKA, P. and YE, J. (2012). Tensor completion for estimating missing values in visual data. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 208–220.

MAO, Y. and SAUL, L. K. (2004). Modeling distances in large-scale networks by matrix factorization. In *Proceedings of the* 4*th ACM SIGCOMM Conference on Internet Measurement* 278–287.

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

MENDILLO, M. (2006). Storms in the ionosphere: Patterns and processes for total electron content. *Reviews of Geophysics* **44** RG4001.

PAATERO, P. and TAPPER, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5** 111–126.

PRÖLSS, G. W. (2008). Ionospheric storms at mid-latitude: A short review. In *Geophysical Monograph Series* (P. M. Kintner, A. J. Coster, T. Fuller-Rowell, A. J. Mannucci, M. Mendillo and R. Heelis, eds.) **181** 9–24. American Geophysical Union, Washington, DC.

RIDEOUT, W. and COSTER, A. (2006). Automated GPS processing for global total electron content data. *GPS Solutions* **10** 219–228.

ROMA-DOLLASE, D., HERNÁNDEZ-PAJARES, M., KRANKOWSKI, A., KOTULAK, K., GHODDOUSI-FARD, R., YUAN, Y., LI, Z., ZHANG, H., SHI, C. et al. (2018). Consistency of seven different GNSS global ionospheric mapping techniques during one solar cycle. *Journal of Geodesy* **92** 691–706.

SCHAER, S. (1999). Mapping and predicting the Earth's ionosphere using the global positioning system. *Dissertation of Astronomical Institute* 205.

SCHAER, S., BEUTLER, G., MERVART, L., ROTHACHER, M. and WILD, U. (1995). Global and regional ionosphere models using the GPS double difference phase observable. *Proceeding of the IGS Workshop on Special Topics and New Directions* 77–92.

SHEPHERD, S. G. (2014). Altitude-adjusted corrected geomagnetic coordinates: Definition and functional approximations. *J. Geophys. Res.* **119** 7501–7521.

SREBRO, N., RENNIE, J. and JAAKKOLA, T. S. (2005). Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems* 1329–1336.

SUN, H., HUA, Z., REN, J., ZOU, S., SUN, Y. and CHEN, Y. (2022). Supplement to "'Matrix completion methods for the total electron content video reconstruction." https://doi.org/10.1214/21-AOAS1541SUPPA, https://doi.org/10.1214/21-AOAS1541SUPPB

THE WHITE HOUSE (2019). National space weather strategy and action plan.

VIERINEN, J., COSTER, A. J., RIDEOUT, W. C., ERICKSON, P. J. and NORBERG, J. (2016). Statistical framework for estimating GNSS bias. *Atmos. Meas. Tech.* **9** 1303–1312.

WANG, H., NIE, F. and HUANG, H. (2014). Low-rank tensor completion with spatio-temporal consistency. In *AAAI* 2846–2852.

XU, Y. and YIN, W. (2013). A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM J. Imaging Sci.* **6** 1758–1789. MR3105787 https://doi.org/10.1137/120887795

XU, Y., HAO, R., YIN, W. and SU, Z. (2015). Parallel matrix factorization for low-rank tensor completion. *Inverse Probl. Imaging* **9** 601–624. MR3356574 https://doi.org/10.3934/ipi.2015.9.601

YANG, Z., MORTON, Y. J., ZAKHARENKOVA, I., CHERNIAK, I., SONG, S. and LI, W. (2020). Global view of ionospheric disturbance impacts on kinematic GPS positioning solutions during the 2015 st. Patrick's day storm. *J. Geophys. Res.* **125** e2019JA027681.

ZHANG, H., XU, P., HAN, W., GE, M. and SHI, C. (2013). Eliminating negative VTEC in global ionosphere maps using inequality-constrained least squares. *Advances in Space Research* **51** 988–1000.

ZOU, S. and RIDLEY, A. J. (2016). Modeling of the evolution of storm-enhanced density plume during the 24 to 25 October 2011 geomagnetic storm. *Magnetosphere-ionosphere Coupling in the Solar System* **10** 205–213.

ZOU, S., RIDLEY, A. J., MOLDWIN, M. B., NICOLLS, M. J., COSTER, A. J., THOMAS, E. G. and RUOHONIEMI, J. M. (2013a). Multi-instrument observations of SED during 24-25 October 2011 storm: Implications for SED formation processes: sed formation processes. *J. Geophys. Res.* **118** 7798–7809.

ZOU, S., MOLDWIN, M. B., NICOLLS, M. J., RIDLEY, A. J., COSTER, A. J., YIZENGAW, E., LYONS, L. R. and DONOVAN, E. F. (2013b). Electrodynamics of the high-latitude trough: Its relationship with convection flows and field-aligned currents. *J. Geophys. Res.* **118** 2565–2572.

ZOU, S., MOLDWIN, M. B., RIDLEY, A. J., NICOLLS, M. J., COSTER, A. J., THOMAS, E. G. and RUOHONIEMI, J. M. (2014). On the generation/decay of the storm-enhanced density plumes: Role of the convection flow and field-aligned ion flow: Generation and decay of SED plumes. *J. Geophys. Res.* **119** 8543–8559.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1359–1379
https://doi.org/10.1214/21-AOAS1545

THE CAUSAL EFFECT OF A TIMEOUT AT STOPPING AN OPPOSING RUN IN THE NBA

BY CONNOR P. GIBBS[1,a], RYAN ELMORE[2,c] AND BAILEY K. FOSDICK[1,b]

[1]*Department of Statistics, Colorado State University,* [a]*Connor.Gibbs@colostate.edu,* [b]*Bailey.Fosdick@colostate.edu*
[2]*Department of Business Information and Analytics, Daniels College of Business, University of Denver,* [c]*Ryan.Elmore@du.edu*

In the summer of 2017, the National Basketball Association reduced the number of total timeouts, along with other rule changes, to regulate the flow of the game. With these rule changes it becomes increasingly important for coaches to effectively manage their timeouts. Understanding the utility of a timeout under various game scenarios, for example, during an opposing team's run, is of the utmost importance. There are two schools of thought when the opposition is on a run: (1) call a timeout and allow your team to rest and regroup, or (2) save a timeout and hope your team can make corrections during play. This paper investigates the credence of these tenets using the Rubin causal model framework to quantify the causal effect of a timeout in the presence of an opposing team's run. Too often overlooked, we carefully consider the stable unit-treatment-value assumption (SUTVA) in this context and use the SUTVA to motivate our definition of units. To measure the effect of a timeout, we introduce a novel, interpretable outcome based on the score difference to describe broad changes in the scoring dynamics. This outcome is well suited for situations where the quantity of interest fluctuates frequently, a commonality in many sports analytics applications. We conclude from our analysis that, while comebacks frequently occur after a run, it is slightly disadvantageous to call a timeout during a run by the opposing team and further demonstrate that the magnitude of this effect varies by franchise.

REFERENCES

ABADIE, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *J. Amer. Statist. Assoc.* **97** 284–292. MR1947286 https://doi.org/10.1198/016214502753479419

ABADIE, A., DRUKKER, D., HERR, J. L. and IMBENS, G. W. (2004). Implementing matching estimators for average treatment effects in Stata. *Stata J.* **4** 290–311.

ASCHBURNER, S. (2017). NBA changes timeout rules to improve game flow. Available at https://www.nba.com/article/2017/07/12/nba-board-governors-timeout-rules-game-flow-trade-deadline [Accessed: 2020-11-12].

ASSIS, N., ASSUNÇÃO, R. and VAZ-DE-MELO, P. O. (2020). Stop the clock: Are timeout effects real? Preprint. Available at arXiv:2009.06750.

AVUGOS, S., KÖPPEN, J., CZIENSKOWSKI, U., RAAB, M. and BAR-ELI, M. (2013). The "hot hand" reconsidered: A meta-analytic approach. *Psychol. Sport Exerc.* **14** 21–27.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BEUOY, M. (2021). Inpredictable: NBA per possession statistics. Available at http://stats.inpredictable.com/nba/ssnTeamPoss.php?season=2018&po=0&frdt=2017-10-17&todt=2019-06-13&view=off [Accessed: 2020-11-12].

BOREN, C. (2019). Raptors' coach Nick Nurse ripped for timeout that helped the Warriors. Available at https://www.washingtonpost.com/sports/2019/06/11/raptors-coach-nick-nurse-ripped-timeout-that-helped-warriors/ [Accessed: 2020-11-12].

BRESLER, A. (2019). nbastatR: R's interface to NBA data. R package version 0.1.120301.

CURTIS, C. (2019). The NBA rule about timeouts that partially explains Nick Nurse's weird decision to take one. Available at https://ftw.usatoday.com/2019/06/nba-finals-raptors-nick-nurse-timeout-rule [Accessed: 2020-11-12].

Key words and phrases. Causal inference, matching, sports statistics.

DESHPANDE, S. K. and EVANS, K. (2020). Expected hypothetical completion probability. *J. Quant. Anal. Sports* **16** 85–94.
DIAMOND, A. and SEKHON, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Rev. Econ. Stat.* **95** 932–945.
ESPN (2019). Nick Nurse disrupted Kawhi's flow, late timeout cost the Raptors in Game 5. Available at https://www.youtube.com/watch?v=UvOvNy-3etc [Accessed: 2020-11-12].
FRANKS, A., MILLER, A., BORNN, L. and GOLDSBERRY, K. (2015). Characterizing the spatial structure of defensive skill in professional basketball. *Ann. Appl. Stat.* **9** 94–121. MR3341109 https://doi.org/10.1214/14-AOAS799
GIBBS, C. P., ELMORE, R. and FOSDICK, B. K. (2022a). Supplement A to "The causal effect of a timeout at stopping an opposing run in the NBA." https://doi.org/10.1214/21-AOAS1545SUPPA
GIBBS, C. P., ELMORE, R. and FOSDICK, B. K. (2022b). Supplement B to "The causal effect of a timeout at stopping an opposing run in the NBA." https://doi.org/10.1214/21-AOAS1545SUPPB
GILOVICH, T., VALLONE, R. and TVERSKY, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cogn. Psychol.* **17** 295–314.
HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. CRC Press, London. MR1082147
HECKMAN, J. J., ICHIMURA, H. and TODD, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Rev. Econ. Stud.* **64** 605–654.
HECKMAN, J. J., ICHIMURA, H. and TODD, P. (1998). Matching as an econometric evaluation estimator. *Rev. Econ. Stud.* **65** 261–294. MR1623713 https://doi.org/10.1111/1467-937X.00044
IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—For Statistics, Social, and Biomedical Sciences: An introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751
KOEHLER, J. J. and CONLEY, C. A. (2003). The "hot hand" myth in professional basketball. *J. Sport Exerc. Psychol.* **25** 253–259.
LAULETTA, T. (2019). Nick Nurse called a bizarre timeout late in the fourth quarter that killed the Raptors momentum and sent the Warriors on their game-winning hot streak. Available at https://www.businessinsider.com/nick-nurse-timeout-warriors-hot-streak-raptors-momentum-2019-6/ [Accessed: 2020-11-12].
LIU, W., KURAMOTO, S. J. and STUART, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in non-experimental prevention research. *Prev. Sci.* **14** 570–580.
LOPEZ, M. J. and GUTMAN, R. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. *Statist. Sci.* **32** 432–454. MR3696004 https://doi.org/10.1214/17-STS612
LOPEZ, M. J., MATTHEWS, G. J. and BAUMER, B. S. (2018). How often does the best team win? A unified approach to understanding randomness in North American sport. *Ann. Appl. Stat.* **12** 2483–2516. MR3875709 https://doi.org/10.1214/18-AOAS1165
MILLER, J. B. and SANJURJO, A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica* **86** 2019–2047. MR3901245 https://doi.org/10.3982/ECTA14943
NBA (2020). NBA advanced stats. Available at https://stats.nba.com [Accessed: 2020-11-12].
PERMUTT, S. (2011). The efficacy of momentum-stopping timeouts on short-term performance in the National Basketball Association. Ph.D. thesis.
PINA, M. (2019). To call a timeout, or not to call a timeout: That is the question for NBA coaches. Available at https://www.si.com/nba/2019/05/02/steve-kerr-gregg-popovich-mike-dantoni-brad-stevens-brett-brown-nba-coaches-playoffs-timeouts [Accessed: 2020-11-12].
R CORE TEAM (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
ROSENBAUM, P. R. (2007). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics* **63** 456–464. MR2370804 https://doi.org/10.1111/j.1541-0420.2006.00717.x
ROSENBAUM, P. R. (2013). Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics* **69** 118–127. MR3058058 https://doi.org/10.1111/j.1541-0420.2012.01821.x
ROSENBAUM, P. R. (2015). Two R packages for sensitivity analysis in observational studies. *Obs. Stud.* **1** 1–17.
ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41
ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.
RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.
RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581
RUBIN, D. B. (1977). Assignment to treatment group on the basis of a covariate. *J. Educ. Stat.* **2** 1–26.

SAAVEDRA, S., MUKHERJEE, S. and BAGROW, J. P. (2012). Is coaching experience associated with effective use of timeouts in basketball? *Sci. Rep.* **2** 676.

SEKHON, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *J. Stat. Softw.* **42** 1–52.

SMITH, J. A. and TODD, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *J. Econometrics* **125** 305–353. MR2143379 https://doi.org/10.1016/j.jeconom.2004.04.011

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313

TOUMI, A. and LOPEZ, M. (2019). From grapes and prunes to apples and apples: Using matched methods to estimate optimal zone entry decision-making in the National Hockey League. In *Carnegie Mellon Sports Analytics Conference*.

VOCK, D. M. and VOCK, L. F. B. (2018). Estimating the effect of plate discipline using a causal inference framework: An application of the G-computation algorithm. *J. Quant. Anal. Sports* **14** 37–56.

YAM, D. R. and LOPEZ, M. J. (2019). What was lost? A causal estimate of fourth down behavior in the National Football League. *J. Sports Anal.* **5** 153–167.

YOUSUF, S. (2018). Rick Carlisle is one of the NBA's best tacticians. How does he choose when to call a timeout? 'It's not a simple answer.'. Available at https://theathletic.com/723304/2018/12/19/rick-carlisle-is-one-of-the-nbas-best-taciticians-how-does-he-choose-when-to-call-a-timeout-its-not-a-simple-answer [Accessed: 2020-11-12].

ZHANG, Z., KIM, H. J., LONJON, G. and ZHU, Y. (2019). Balance diagnostics after propensity score matching. *Ann. Transl. Med.* **7** 16.

ZIMMERMAN, D. L., TANG, J. and HUANG, R. (2019). Outline analyses of the called strike zone in Major League Baseball. *Ann. Appl. Stat.* **13** 2416–2451. MR4037436 https://doi.org/10.1214/19-aoas1285

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1380–1399
https://doi.org/10.1214/21-AOAS1546

BAYESIAN SEMIPARAMETRIC LONG MEMORY MODELS FOR DISCRETIZED EVENT DATA

BY ANTIK CHAKRABORTY[1,a], OTSO OVASKAINEN[3,4,5,c] AND DAVID B. DUNSON[2,b]

[1]*Department of Statistics, Purdue University,* [a]*antik015@purdue.edu*
[2]*Department of Statistical Science, Duke University,* [b]*dunson@duke.edu*
[3]*Department of Biological and Environmental Science, University of Jyväskylä,* [c]*otso.t.ovaskainen@jyu.fi*
[4]*Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki*
[5]*Department of Biology, Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim*

We introduce a new class of semiparametric latent variable models for long memory discretized event data. The proposed methodology is motivated by a study of bird vocalizations in the Amazon rain forest; the timings of vocalizations exhibit self-similarity and long range dependence. This rules out Poisson process based models where the rate function itself is not long range dependent. The proposed class of FRActional Probit (FRAP) models is based on thresholding, a latent process. This latent process is modeled by a smooth Gaussian process and a fractional Brownian motion by assuming an additive structure. We develop a Bayesian approach to inference using Markov chain Monte Carlo and show good performance in simulation studies. Applying the methods to the Amazon bird vocalization data, we find substantial evidence for self-similarity and non-Markovian/Poisson dynamics. To accommodate the bird vocalization data in which there are many different species of birds exhibiting their own vocalization dynamics, a hierarchical expansion of FRAP is provided in the Supplementary Material.

REFERENCES

BERAN, J., FENG, Y., GHOSH, S. and KULIK, R. (2013). *Long-Memory Processes*: *Probabilistic Properties and Statistical Methods*. Springer, Heidelberg. MR3075595 https://doi.org/10.1007/978-3-642-35512-7

CHAKRABORTY, A., OVASKAINEN, O. and DUNSON, D. B (2022a). Supplement to "Bayesian semiparametric long memory models for discretized event data." https://doi.org/10.1214/21-AOAS1546SUPPA

CHAKRABORTY, A., OVASKAINEN, O. and DUNSON, D. B (2022b). Code to implement methods in "Bayesian semiparametric long memory models for discretized event data." https://doi.org/10.1214/21-AOAS1546SUPPB

CHEN, Y., HÄRDLE, W. K. and PIGORSCH, U. (2010). Localized realized volatility modeling. *J. Amer. Statist. Assoc*. **105** 1376–1393. MR2796557 https://doi.org/10.1198/jasa.2010.ap09039

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

CUI, Y. and LUND, R. (2009). A new look at time series of counts. *Biometrika* **96** 781–792. MR2564490 https://doi.org/10.1093/biomet/asp057

CZADO, C., GNEITING, T. and HELD, L. (2009). Predictive model assessment for count data. *Biometrics* **65** 1254–1261. MR2756513 https://doi.org/10.1111/j.1541-0420.2009.01191.x

DAVIS, R. A., HOLAN, S. H., LUND, R. and RAVISHANKER, N. (2016). *Handbook of Discrete-Valued Time Series*. CRC Press.

DAVISON, A. and RAMESH, N. (1996). Some models for discretized series of events. *J. Amer. Statist. Assoc*. **91** 601–609.

DE CAMARGO, U., ROSLIN, T. and OVASKAINEN, O. (2019). Spatio-temporal scaling of biodiversity in acoustic tropical bird communities. *Ecography* **42** 1936–1947.

FEARNHEAD, P. and SHERLOCK, C. (2006). An exact Gibbs sampler for the Markov-modulated Poisson process. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **68** 767–784. MR2301294 https://doi.org/10.1111/j.1467-9868.2006.00566.x

Key words and phrases. Fractional Brownian motion, fractal, latent Gaussian process models, long range dependence, nonparametric Bayes, probit, time series.

FISCHER, W. and MEIER-HELLSTERN, K. (1993). The Markov-modulated Poisson process (MMPP) cookbook. *Perform. Eval.* **18** 149–171. MR1237373 https://doi.org/10.1016/0166-5316(93)90035-S

FRANZKE, C. L., BARBOSA, S., BLENDER, R., FREDRIKSEN, H.-B., LAEPPLE, T., LAMBERT, F., NILSEN, T., RYPDAL, K., RYPDAL, M. et al. (2020). The structure of climate variability across scales. *Reviews of Geophysics* **58** e2019RG000657.

GEWEKE, J. and PORTER-HUDAK, S. (1983). The estimation and application of long memory time series models. *J. Time Series Anal.* **4** 221–238. MR0738585 https://doi.org/10.1111/j.1467-9892.1983.tb00371.x

GRAVES, T., GRAMACY, R., WATKINS, N. and FRANZKE, C. (2017). A brief history of long memory: Hurst, Mandelbrot and the road to ARFIMA, 1951–1980. *Entropy* **19** 437.

HALL, P. and HART, J. D. (1990). Nonparametric regression with long-range dependence. *Stochastic Process. Appl.* **36** 339–351. MR1084984 https://doi.org/10.1016/0304-4149(90)90100-7

HURST, H. E. (1951). Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civ. Eng.* **116** 770–799.

JACOBS, P. A. and LEWIS, P. A. W. (1978a). Discrete time series generated by mixtures. I. Correlational and runs properties. *J. Roy. Statist. Soc. Ser. B* **40** 94–105. MR0512147

JACOBS, P. A. and LEWIS, P. A. W. (1978b). Discrete time series generated by mixtures. II. Asymptotic properties. *J. Roy. Statist. Soc. Ser. B* **40** 222–228. MR0517443

JIA, Y., KECHAGIAS, S., LIVSEY, J., LUND, R. and PIPIRAS, V. (2021). Latent Gaussian count time series. *J. Amer. Statist. Assoc.* 1–28.

KOLASSA, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *Int. J. Forecast.* **32** 788–803.

KREBS, J. R. and KACELNIK, A. (1983). The dawn chorus in the great tit (Parus major): Proximate and ultimate causes. *Behaviour* **83** 287–308.

LAIOLO, P. (2010). The emerging significance of bioacoustics in animal species conservation. *Biol. Conserv.* **143** 1635–1645.

LASKIN, N. (2003). Fractional Poisson process *Commun. Nonlinear Sci. Numer. Simul.* **8** 201–213. MR2007003 https://doi.org/10.1016/S1007-5704(03)00037-6

LIVSEY, J., LUND, R., KECHAGIAS, S. and PIPIRAS, V. (2018). Multivariate integer-valued time series with flexible autocovariances and their application to major hurricane counts. *Ann. Appl. Stat.* **12** 408–431. MR3773399 https://doi.org/10.1214/17-AOAS1098

LO, A. W. (1989). Long-term memory in stock market prices. Technical Report, National Bureau of Economic Research.

MANDELBROT, B. B. and VAN NESS, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Rev.* **10** 422–437. MR0242239 https://doi.org/10.1137/1010093

MANDELBROT, B. B. and WALLIS, J. R. (1969). Some long-run properties of geophysical records. *Water Resour. Res.* **5** 321–340.

MCKENZIE, E. (1985). Some simple models for discrete variate time series 1. *J. Am. Water Resour. Assoc.* **21** 645–650.

MCKENZIE, E. (1986). Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Adv. in Appl. Probab.* **18** 679–705. MR0857325 https://doi.org/10.2307/1427183

MCKENZIE, E. (1988). Some ARMA models for dependent sequences of Poisson counts. *Adv. in Appl. Probab.* **20** 822–835. MR0968000 https://doi.org/10.2307/1427362

MIKOSCH, T. and STĂRICĂ, C. (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Rev. Econ. Stat.* **86** 378–390.

MITRINOVIC, D. S. and VASIC, P. M. (1970). *Analytic Inequalities* **1**. Springer.

OGATA, Y. and ABE, K. (1991). Some statistical features of the long-term variation of the global and regional seismic activity. *International Statistical Review/Revue Internationale de Statistique* 139–161.

OVASKAINEN, O., DE CAMARGO, U. M. and SOMERVUO, P. (2018). Animal sound identifier (ASI): Software for automated identification of vocal animals. *Ecol. Lett.* **21** 1244–1254. https://doi.org/10.1111/ele.13092

PENG, C.-K., BULDYREV, S. V., HAVLIN, S., SIMONS, M., STANLEY, H. E. and GOLDBERGER, A. L. (1994). Mosaic organization of DNA nucleotides. *Phys. Rev. E* **49** 1685.

PIPIRAS, V. and TAQQU, M. S. (2017). *Long-Range Dependence and Self-Similarity. Cambridge Series in Statistical and Probabilistic Mathematics* **45**. Cambridge Univ. Press, Cambridge. MR3729426

RAMESH, N., THAYAKARAN, R. and ONOF, C. (2013). Multi-site doubly stochastic Poisson process models for fine-scale rainfall. *Stoch. Environ. Res. Risk Assess.* **27** 1383–1396.

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

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

ROBINSON, P. M. (1995). Gaussian semiparametric estimation of long range dependence. *Ann. Statist.* **23** 1630–1661. MR1370301 https://doi.org/10.1214/aos/1176324317

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

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

SAMORODNITSKY, G. (2006). Long range dependence. *Found. Trends Stoch. Syst.* **1** 163–257. MR2379935 https://doi.org/10.1561/0900000004

SLABBEKOORN, H. and SMITH, T. B. (2002). Bird song, ecology and speciation. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **357** 493–503. https://doi.org/10.1098/rstb.2001.1056

SØRBYE, S. H. and RUE, H. (2018). Fractional Gaussian noise: Prior specification and model comparison. *Environmetrics* **29** e2457. MR3830875 https://doi.org/10.1002/env.2457

STERN, R. and COE, R. (1984). A model fitting analysis of daily rainfall data. *J. R. Stat. Soc., A* **147** 1–18.

TAGLIAZUCCHI, E., VON WEGNER, F., MORZELEWSKI, A., BRODBECK, V., JAHNKE, K. and LAUFS, H. (2013). Breakdown of long-range temporal dependence in default mode and attention networks during deep sleep. *Proc. Natl. Acad. Sci. USA* **110** 15419–15424.

TIAO, G. C., PHADKE, M. and BOX, G. E. (1976). Some empirical models for the Los Angeles photochemical smog data. *J. Air Pollut. Control Assoc.* **26** 485–490.

WEYMER, B. A., WERNETTE, P., EVERETT, M. E. and HOUSER, C. (2018). Statistical modeling of the long-range dependent structure of barrier island framework geology and surface geomorphology. *Earth Surf. Dyn.* **6** 431–450.

WILLINGER, W., PAXSON, V., RIEDI, R. H. and TAQQU, M. S. (2003). Long-range dependence and data network traffic. In *Theory and Applications of Long-Range Dependence* 373–407. Birkhäuser, Boston, MA. MR1957500

ZHOU, Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. *J. Time Series Anal.* **33** 438–457. MR2915095 https://doi.org/10.1111/j.1467-9892.2011.00780.x

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1400–1422
https://doi.org/10.1214/21-AOAS1547

CO-CLUSTERING OF MULTIVARIATE FUNCTIONAL DATA FOR THE ANALYSIS OF AIR POLLUTION IN THE SOUTH OF FRANCE

BY CHARLES BOUVEYRON[1,a], JULIEN JACQUES[2,b], AMANDINE SCHMUTZ[2,c], FANNY SIMÕES[3,d] AND SILVIA BOTTINI[3,e]

[1]*Université Côte d'Azur, Inria, CNRS, LJAD, Maasai,* [a]*charles.bouveyron@univ-cotedazur.fr*
[2]*Université de Lyon, Lyon 2, Laboratoire ERIC,* [b]*julien.jacques@univ-lyon2.fr,* [c]*schmutz.amandine@gmail.com*
[3]*Université Côte d'Azur, MSI,* [d]*fanny.simoes@univ-cotedazur.fr,* [e]*silvia.bottini@univ-cotedazur.fr*

Nowadays, air pollution is a major threat for public health with clear relationships with many diseases, especially cardiovascular ones. The spatiotemporal study of pollution is of great interest for governments and local authorities when deciding for public alerts or new city policies against pollution increase. The aim of this work is to study spatiotemporal profiles of environmental data collected in the south of France (Région Sud) by the public agency AtmoSud. The idea is to better understand the exposition to pollutants of inhabitants on a large territory with important differences in term of geography and urbanism. The data gather the recording of daily measurements of five environmental variables, namely, three pollutants (PM10, NO_2, O_3) and two meteorological factors (pressure and temperature) over six years. Those data can be seen as multivariate functional data: quantitative entities evolving along time for which there is a growing need of methods to summarize and understand them. For this purpose a novel co-clustering model for multivariate functional data is defined. The model is based on a functional latent block model which assumes for each co-cluster a probabilistic distribution for multivariate functional principal component scores. A stochastic EM algorithm, embedding a Gibbs sampler, is proposed for model inference as well as a model selection criteria for choosing the number of co-clusters. The application of the proposed co-clustering algorithm on environmental data of the Région Sud allowed to divide the region, composed by 357 zones, into six macroareas with common exposure to pollution. We showed that pollution profiles vary accordingly to the seasons, and the patterns are similar during the six years studied. These results can be used by local authorities to develop specific programs to reduce pollution at the macroarea level and to identify specific periods of the year with high pollution peaks in order to set up specific health prevention programs. Overall, the proposed co-clustering approach is a powerful resource to analyse multivariate functional data in order to identify intrinsic data structure and to summarize variables profiles over long periods of time.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **9** 716–723. MR0423716 https://doi.org/10.1109/tac.1974.1100705

BANERJEE, A., DHILLON, I., GHOSH, J., MERUGU, S. and MODHA, D. S. (2007). A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. *J. Mach. Learn. Res.* **8** 1919–1986. MR2353824 https://doi.org/10.1145/1014052.1014111

BEN SLIMEN, Y., ALLIO, S. and JACQUES, J. (2018). Model-based co-clustering for functional data. *Neurocomputing* **291** 97–108.

BENBRAHIM-TALLAA, L., BAAN, R., GROSSE, Y., LAUBY-SECRETAN, B., EL GHISSASSI, F. and BOUVARD, V. E. A. (2012). Carcinogenicity of diesel-engine and gasoline-engine exhausts and some nitroarenes. *Lancet Oncol.* **13** 663–664.

Key words and phrases. Latent block model, multivariate functional data, SEM-Gibbs algorithm, pollution, co-clustering.

BHATIA, P., IOVLEFF, S. and GOVAERT, G. (2017). blockcluster: An R package for model based co-clustering. *J. Stat. Softw.* **9** 1–24.

BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 719–725.

BOUVEYRON, C., CÔME, E. and JACQUES, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *Ann. Appl. Stat.* **9** 1726–1760. MR3456352 https://doi.org/10.1214/15-AOAS861

BOUVEYRON, C., JACQUES, J. and SCHMUTZ, A. (2020). funLBM: Model-Based Co-Clustering of Functional Data. R package version 2.1.

BOUVEYRON, C., BOZZI, L., JACQUES, J. and JOLLOIS, F.-X. (2018). The functional latent block model for the co-clustering of electricity consumption curves. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 897–915. MR3832256 https://doi.org/10.1111/rssc.12260

BOUVEYRON, C., CELEUX, G., MURPHY, T. B. and RAFTERY, A. E. (2019). *Model-Based Clustering and Classification for Data Science*: *With applications in R. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, Cambridge. MR3967046 https://doi.org/10.1017/9781108644181

BOUVEYRON, C., JACQUES, J., SCHMUTZ, A., SIMÕES, F. and BOTTINI, S. (2022a). Supplement to "Co-Clustering of multivariate functional data for the analysis of air pollution in the South of France." https://doi.org/10.1214/21-AOAS1547SUPPA

BOUVEYRON, C., JACQUES, J., SCHMUTZ, A., SIMÕES, F. and BOTTINI, S. (2022b). Supplement (data and code) to "Co-Clustering of multivariate functional data for the analysis of air pollution in the South of France." https://doi.org/10.1214/21-AOAS1547SUPPB

CHAMROUKHI, F. and BIERNACKI, C. (2017). Model-based co-clustering of multivariate functional data. In *ISI* 2017—61*st World Statistics Congress*, Marrakech, Morocco.

CORNELI, M., BOUVEYRON, C. and LATOUCHE, P. (2020). Co-clustering of ordinal data via latent continuous random variables and not missing at random entries. *J. Comput. Graph. Statist.* **29** 771–785. MR4191242 https://doi.org/10.1080/10618600.2020.1739533

DELAIGLE, A. and HALL, P. (2010). Defining probability density for a distribution of random functions. *Ann. Statist.* **38** 1171–1193. MR2604709 https://doi.org/10.1214/09-AOS741

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DI ORIO, J. and VANTINI, S. (2019). funBI: A biclustering algorithm for functional data.

GEORGE, T. and MERUGU, S. (2005). A scalable collaborative filtering framework based on co-clustering. In *Data Mining*, *Fifth IEEE International Conference on*. IEEE, New York.

GOVAERT, G. and NADIF, M. (2013). *Co-Clustering*, 1st ed. Wiley-IEEE Press, New York.

HAMRA, G., GUHA, N., COHEN, A., LADEN, F., RAASCHOU-NIELSEN, O., SAMET, J. et al. (2014). Outdoor particulate matter exposure and lung cancer: A systematic review and meta-analysis. *Environ. Health Perspect.* **112** 906–911.

IARC (2016). Outdoor air pollution. Volume 109 of *IARC Monogr. Eval. Carcinog. Risks. Hum.*

IEVA, F., PAGANONI, A. M., PIGOLI, D. and VITELLI, V. (2013). Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 401–418. MR3060623 https://doi.org/10.1111/j.1467-9876.2012.01062.x

JACQUES, J. and BIERNACKI, C. (2018). Model-based co-clustering for ordinal data. *Comput. Statist. Data Anal.* **123** 101–115. MR3777088 https://doi.org/10.1016/j.csda.2018.01.014

JACQUES, J. and PREDA, C. (2013). Funclust: A curves clustering method using functional random variable density approximation. *Neurocomputing* **112** 164–171.

JACQUES, J. and PREDA, C. (2014a). Functional data clustering: A survey. *Adv. Data Anal. Classif.* **8** 231–255. MR3253859 https://doi.org/10.1007/s11634-013-0158-y

JACQUES, J. and PREDA, C. (2014b). Model-based clustering for multivariate functional data. *Comput. Statist. Data Anal.* **71** 92–106. MR3131956 https://doi.org/10.1016/j.csda.2012.12.004

KAYANO, M., DOZONO, K. and KONISHI, S. (2010). Functional cluster analysis via orthonormalized Gaussian basis expansions and its application. *J. Classification* **27** 211–230. MR2726319 https://doi.org/10.1007/s00357-010-9054-8

KERIBIN, C., GOVAERT, G. and CELEUX, G. (2010). Estimation d'un modèle à blocs latents par l'algorithme SEM. In 42*èmes Journées de Statistique*, Marseille, France, France.

KERIBIN, C., BRAULT, V., CELEUX, G. and GOVAERT, G. (2015). Estimation and selection for the latent block model on categorical data. *Stat. Comput.* **25** 1201–1216. MR3401881 https://doi.org/10.1007/s11222-014-9472-2

LACLAU, C., REDKO, I., MATEI, B., BENNANI, Y. and BRAULT, V. (2017). Co-clustering through optimal transport. In 34*th International Conference on Machine Learning*. *Proceedings of the* 34*th International Conference on Machine Learning* **70** 1955–1964. Proceedings of Machine Learning Research, Sydney, Australia.

LELIEVELD, J., EVANS, J. and FNAIS, M. E. A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature* **525** 367–371.

MARTÍNEZ-HERNÁNDEZ, I. and GENTON, M. G. (2020). Recent developments in complex and spatially correlated functional data. *Braz. J. Probab. Stat.* **34** 204–229. MR4093256 https://doi.org/10.1214/20-BJPS466

MENUT, L., BESSAGNET, B., KHVOROSTYANOV, D., BEEKMANN, M., BLOND, N., COLETTE, A., COLL, I., CURCI, G., FORET, G. et al. (2013). Chimere 2013: A model for regional atmospheric composition modelling. *Geosci. Model Dev.* **6** 981–1028.

NADIF, M. and GOVAERT, G. (2008). Algorithms for model-based block Gaussian clustering. In *Proceedings of the* 2008 *International Conference on Data Mining*, *DMIN* 2008, *July* 14–17, 2008, 2 *Volumes* 536–542, Las Vegas, USA.

PASCAL, M., DE CROUY CHANEL, P., WAGNER, V., CORSO, M., TILLIER, C., BENTAYEB, M., BLANCHARD, M., COCHET, A., PASCAL, L. et al. (2016). The mortality impacts of fine particles in France. *Sci. Total Environ.* **571** 416–425.

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

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.

SCHMUTZ, A., JACQUES, J., BOUVEYRON, C., CHÉZE, L. and MARTIN, P. (2020). Clustering multivariate functional data in group-specific functional subspaces. *Comput. Statist.* **35** 1101–1131. MR4133110 https://doi.org/10.1007/s00180-020-00958-4

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SELOSSE, M., JACQUES, J. and BIERNACKI, C. (2020). Model-based co-clustering for mixed type data. *Comput. Statist. Data Anal.* **144** 106866. MR4023907 https://doi.org/10.1016/j.csda.2019.106866

TOKUSHIGE, S., YADOHISA, H. and INADA, K. (2007). Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Comput. Statist.* **22** 1–16. MR2299243 https://doi.org/10.1007/s00180-006-0013-0

VANDEWALLE, V., PREDA, C. and DABO-NIANG, S. (2020). Clustering spatial functional data. In *Geostatistical Functional Data Analysis*: *Theory and Methods* (J. Mateu and R. Giraldo, eds.) Wiley, Chichester.

WANG, S. and HUANG, A. (2017). Penalized nonnegative matrix tri-factorization for co-clustering. *Expert Syst. Appl.* **78** 64–73.

WHO REGIONAL OFFICE FOR EUROPE (2013). Review of evidence on health aspects of air pollution—REVIHAAP Project. Technical report, Copenhagen, Denmark.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1423–1444
https://doi.org/10.1214/21-AOAS1548

INTEGRATED QUANTILE RANK TEST (IQRAT) FOR GENE-LEVEL ASSOCIATIONS

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

[1]*Center for Statistical Science & Department of Industrial Engineering, Tsinghua University,* [a]*tianyingw@tsinghua.edu.cn*
[2]*Department of Biostatistics, Columbia University,* [b]*ii2135@cumc.columbia.edu,* [c]*yw2148@cumc.columbia.edu*

Gene-based testing is a commonly employed strategy in many genetic association studies. Gene-trait associations can be complex due to underlying population heterogeneity, gene-environment interactions, and various other reasons. Existing gene-based tests, such as burden and sequence kernel association tests (SKAT), are mean-based tests and may miss or underestimate higher-order associations that could be scientifically interesting. In this paper we propose a new family of gene-level association tests that integrate quantile rank score process to better accommodate complex associations. The resulting test statistics have multiple advantages: (1) they are almost as efficient as the best existing tests when the associations are homogeneous across quantile levels and have improved efficiency for complex and heterogeneous associations; (2) they provide useful insights into risk stratification; (3) the test statistics are distribution free and could hence accommodate a wide range of underlying distributions, and (4) they are computationally efficient. We established the asymptotic properties of the proposed tests under the null and alternative hypotheses and conducted large-scale simulation studies to investigate their finite sample performance. The performance of the proposed approach is compared with that of conventional mean-based tests, that is, the burden and SKAT tests, through simulation studies and applications to a metabochip dataset on lipid traits and to the genotype-tissue expression data in GTEx to identify eGenes, that is, genes whose expression levels are associated with cis-eQTLs.

REFERENCES

BACKENROTH, D., HE, Z., KIRYLUK, K., BOEVA, V., PETHUKOVA, L., KHURANA, E., CHRISTIANO, A., BUXBAUM, J. D. and IONITA-LAZA, I. (2018). FUN-LDA: A latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: Methods and applications. *Am. J. Hum. Genet.* **102** 920–942.

BEYERLEIN, A., VON KRIES, R., NESS, A. R. and ONG, K. K. (2011). Genetic markers of obesity risk: Stronger associations with body composition in overweight compared to normal-weight children. *PLoS ONE* **6** e19057. https://doi.org/10.1371/journal.pone.0019057

BOMBA, L., WALTER, K. and SORANZO, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18** 77. https://doi.org/10.1186/s13059-017-1212-4

BRIOLLAIS, L. and DURRIEU, G. (2014). Application of quantile regression to recent genetic and-omic studies. *Hum. Genet.* **133** 951–966.

BROWN, A. A., BUIL, A., VIÑUELA, A., LAPPALAINEN, T., ZHENG, H.-F., RICHARDS, J. B., SMALL, K. S., SPECTOR, T. D., DERMITZAKIS, E. T. et al. (2014). Genetic interactions affecting human gene expression identified by variance association mapping. *eLife* **3** e01381. https://doi.org/10.7554/eLife.01381

CHEN, H., HUFFMAN, J. E., BRODY, J. A., WANG, C., LEE, S., LI, Z., GOGARTEN, S. M., SOFER, T., BIELAK, L. F. et al. (2019). Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *Am. J. Hum. Genet.* **104** 260–274.

DAVIES, R. B. (1980). Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **29** 323–333.

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

Key words and phrases. Quantile process, rank score test, gene-set associations.

FERENCE, B. A., KASTELEIN, J. J., RAY, K. K., GINSBERG, H. N., CHAPMAN, M. J., PACKARD, C. J., LAUFS, U., OLIVER-WILLIAMS, C., WOOD, A. M. et al. (2019). Association of triglyceride-lowering LPL variants and LDL-C–lowering LDLR variants with risk of coronary heart disease. *JAMA* **321** 364–373.

FISHER, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in Statistics* 66–70. Springer, Berlin.

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

GUTENBRUNNER, C. and JUREČKOVÁ, J. (1992). Regression rank scores and regression quantiles. *Ann. Statist.* **20** 305–330. MR1150346 https://doi.org/10.1214/aos/1176348524

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

HÁJEK, J., ŠIDÁK, Z. and SEN, P. K. (1999). *Theory of Rank Tests*, 2nd ed. *Probability and Mathematical Statistics*. Academic Press, San Diego, CA. MR1680991

HAN, F. and PAN, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.* **70** 42–54. https://doi.org/10.1159/000288704

HE, Z., XU, B., LEE, S. and IONITA-LAZA, I. (2017). Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in metabochip data. *Am. J. Hum. Genet.* **101** 340–352.

HE, Z., LIU, L., WANG, K. and IONITA-LAZA, I. (2018). A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using MPRAs. *Nat. Commun.* **9** 5199.

HE, Z., XU, B., BUXBAUM, J. and IONITA-LAZA, I. (2019). A genome-wide scan statistic framework for whole-genome sequence data analysis. *Nat. Commun.* **10** 3018.

HUANG, Y.-F., GULKO, B. and SIEPEL, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49** 618–624. https://doi.org/10.1038/ng.3810

IONITA-LAZA, I., BUXBAUM, J. D., LAIRD, N. M. and LANGE, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.* **7** e1001289. https://doi.org/10.1371/journal.pgen.1001289

IONITA-LAZA, I., LEE, S., MAKAROV, V., BUXBAUM, J. D. and LIN, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* **92** 841–853.

IONITA-LAZA, I., MCCALLUM, K., XU, B. and BUXBAUM, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48** 214.

JENG, X. J., DAYE, Z. J., LU, W. and TZENG, J.-Y. (2016). Rare variants association analysis in large-scale sequencing studies at the single locus level. *PLoS Comput. Biol.* **12** e1004993.

JIN, J. (2006). Higher criticism statistic: Theory and applications in non-Gaussian detection. In *Statistical Problems in Particle Physics, Astrophysics and Cosmology* 233–236. World Scientific, Singapore.

JUSTICE, A. E., HOWARD, A. G., FERNÁNDEZ-RHODES, L., GRAFF, M., TAO, R. and NORTH, K. E. (2018). Direct and indirect genetic effects on triglycerides through omics and correlated phenotypes. *BMC Proc.* **12** 22.

KAI, B., LI, R. and ZOU, H. (2010). Local composite quantile regression smoothing: An efficient and safe alternative to local polynomial regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 49–69. MR2751243 https://doi.org/10.1111/j.1467-9868.2009.00725.x

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

KOENKER, R. (2010). Rank tests for heterogeneous treatment effects with covariates. In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis*: *A Festschrift in Honor of Professor Jana Jurečková*. *Inst. Math. Stat.* (*IMS*) *Collect.* **7** 134–142. IMS, Beachwood, OH. MR2808374

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

KOENKER, R., D'OREY, V. et al. (1990). A note on computing dual regression quantiles and regression rank scores remark on Algorithm 229/BEBR No. 1666. BEBR faculty working paper; no. 90-1666.

KOENKER, R., MIZERA, I. et al. (2014). Convex optimization in R. *J. Stat. Softw.* **60** 1–23.

LEE, S., with contributions from MIROPOLSKY, L. and WU, M. (2017). SKAT: SNP-Set (Sequence) Kernel Association Test. R package version 1.3.2.1. Available at https://CRAN.R-project.org/package=SKAT.

LEE, S., WU, M. C. and LIN, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13** 762–775.

LEE, S., TESLOVICH, T. M., BOEHNKE, M. and LIN, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.* **93** 42–53. PMID: 23768515.

LI, B. and LEAL, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am. J. Hum. Genet.* **83** 311–321.

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

LU, Q., POWLES, R. L., WANG, Q., HE, B. J. and ZHAO, H. (2016). Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLoS Genet.* **12** e1005947.

MADSEN, B. E. and BROWNING, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5** e1000384. https://doi.org/10.1371/journal.pgen.1000384

MANCHIA, M., CULLIS, J., TURECKI, G., ROULEAU, G. A., UHER, R. and ALDA, M. (2013). The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases. *PLoS ONE* **8** e76295. https://doi.org/10.1371/journal.pone.0076295

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

MORRIS, A. P. and ZEGGINI, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* **34** 188–193.

MOSCOVICH, A., NADLER, B. and SPIEGELMAN, C. (2016). On the exact Berk–Jones statistics and their p-value calculation. *Electron. J. Stat.* **10** 2329–2354. MR3544289 https://doi.org/10.1214/16-EJS1172

PARÉ, G., COOK, N. R., RIDKER, P. M. and CHASMAN, D. I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: A report from the Women's Genome Health Study. *PLoS Genet.* **6** e1000981. https://doi.org/10.1371/journal.pgen.1000981

QIU, X., WU, H. and HU, R. (2013). The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC Bioinform.* **14** 124.

QUANG, D., CHEN, Y. and XIE, X. (2014). DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31** 761–763.

SCHAFFNER, S. F., FOO, C., GABRIEL, S., REICH, D., DALY, M. J. and ALTSHULER, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15** 1576–1583.

SCHULTZ, B. B. (1985). Levene's test for relative variation. *Syst. Zool.* **34** 449–456.

SONG, X., LI, G., ZHOU, Z., WANG, X., IONITA-LAZA, I. and WEI, Y. (2017). QRank: A novel quantile regression tool for eQTL discovery. *Bioinformatics* **33** 2123–2130.

SUN, R., HUI, S., BADER, G. D., LIN, X. and KRAFT, P. (2019). Powerful gene set analysis in GWAS with the generalized Berk–Jones statistic. *PLoS Genet.* **15** e1007530.

TALIUN, D., HARRIS, D. N., KESSLER, M. D., CARLSON, J., SZPIECH, Z. A., TORRES, R., GAGLIANO TALIUN, S. A., CORVELO, A., GOGARTEN, S. M. et al. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. BioRxiv 563866.

UEYAMA, C., HORIBE, H., YAMASE, Y., FUJIMAKI, T., OGURI, M., KATO, K., ARAI, M., WATANABE, S., MUROHARA, T. et al. (2015). Association of FURIN and ZPR1 polymorphisms with metabolic syndrome. *Biomed. Reports* **3** 641–647.

VOIGHT, B. F., KANG, H. M., DING, J., PALMER, C. D., SIDORE, C., CHINES, P. S., BURTT, N. P., FUCHSBERGER, C., LI, Y. et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* **8** e1002793. https://doi.org/10.1371/journal.pgen.1002793

WANG, T., IONITA-LAZA, I. and WEI, Y. (2022). Supplement to "Integrated Quantile RAnk Test (iQRAT) for gene-level associations." https://doi.org/10.1214/21-AOAS1548SUPPA, https://doi.org/10.1214/21-AOAS1548SUPPB

WANG, Q., LU, Q. and ZHAO, H. (2015). A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Front. Genet.* **6** 149.

WANG, H., ZHANG, F., ZENG, J., WU, Y., KEMPER, K. E., XUE, A., ZHANG, M., POWELL, J. E., GODDARD, M. E. et al. (2019). Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. bioRxiv 519538.

WEI, Y. and CARROLL, R. J. (2009). Quantile regression with measurement error. *J. Amer. Statist. Assoc.* **104** 1129–1143. MR2562008 https://doi.org/10.1198/jasa.2009.tm08420

WEI, W.-H., HEMANI, G. and HALEY, C. S. (2014). Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **15** 722–733. https://doi.org/10.1038/nrg3747

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

WU, M. C., MAITY, A., LEE, S., SIMMONS, E. M., HARMON, Q. E., LIN, X., ENGEL, S. M., MOLLDREM, J. J. and ARMISTEAD, P. M. (2013). Kernel machine SNP-set testing under multiple candidate kernels. *Genet. Epidemiol.* **37** 267–275.

YANG, J., LOOS, R. J. F., POWELL, J. E., MEDLAND, S. E., SPELIOTES, E. K., CHASMAN, D. I., ROSE, L. M., THORLEIFSSON, G., STEINTHORSDOTTIR, V. et al. (2012). FTO genotype is associated with phenotypic variability of body mass index. *Nature* **490** 267–272.

ZHOU, J. and TROYANSKAYA, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12** 931–934. https://doi.org/10.1038/nmeth.3547

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

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1445–1458
https://doi.org/10.1214/21-AOAS1549

A NOVEL FRAMEWORK TO ESTIMATE MULTIDIMENSIONAL MINIMUM EFFECTIVE DOSES USING ASYMMETRIC POSTERIOR GAIN AND ϵ-TAPERING

BY YING KUEN CHEUNG[1,a], THEVAA CHANDERENG[1,b] AND KEITH M. DIAZ[2,c]

[1]*Department of Biostatistics, Columbia University,* [a]*yc632@cumc.columbia.edu,* [b]*tc3123@columbia.edu*
[2]*Department of Medicine, Columbia University,* [c]*kd2442@columbia.edu*

In this article we address the problem of estimating minimum effective doses in dose-finding clinical trials of multidimensional treatment. We are motivated by a behavioral intervention trial where we introduce sedentary breaks to subjects with a goal to reduce their glucose level monitored over 8 hours. Each sedentary break regimen is defined by two elements: break frequency and break duration. The trial aims to identify minimum combinations of frequency and duration that shift mean glucose, that is, the minimum effective dose (MED) combinations. The means of glucose reduction associated with the dose combinations are only partially ordered. To circumvent constrained estimation due to partial ordering, we propose estimating the MED by maximizing a weighted product of combinationwise posterior gains. The estimation adopts an asymmetric gain function, indexed by a decision parameter ϵ, which defines the relative gains of a true negative decision and a true positive decision. We also introduce an adaptive ϵ-tapering algorithm to be used in conjunction with the estimation method. Simulation studies show that using asymmetric gain with a carefully chosen ϵ is critical to keeping false discoveries low, while ϵ-tapering adds to the probability of identifying truly effective doses (i.e., true positives). Under an ensemble of scenarios for the sedentary break study, ϵ-tapering yields consistently high true positive rates across scenarios and achieves about 90% true positive rate, compared to 68% by a nonadaptive design with comparable false discovery rate.

REFERENCES

ALLISON, D. B., PAULTRE, F., MAGGIO, C., MEZZITIS, N. and PI-SUNYER, F. X. (1995). The use of areas under curves in diabetes research. *Diabetes Care* **18** 245–250.

BISWAS, A., OH, P. I., FAULKNER, G. E., BAJAJ, R. R., SILVER, M. A., MITCHELL, M. S. and ALTER, D. A. (2015). Sedentary time and its association with risk for disease incidence, mortality, and hospitalization in adults: A systematic review and meta-analysis. *Ann. Intern. Med.* **162** 123–132.

BRAUN, T. M. and WANG, S. (2010). A hierarchical Bayesian design for phase 1 trials of novel combinations of cancer therapeutic agents. *Biometrics* **66** 805–812. MR2758216 https://doi.org/10.1111/j.1541-0420.2009.01363.x

CHEUNG, Y. K. (2007). Sequential implementation of stepwise procedures for identifying the maximum tolerated dose. *J. Amer. Statist. Assoc.* **102** 1448–1461. MR2446206 https://doi.org/10.1198/016214507000000699

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

DIAZ, K. M., HOWARD, V. J., HUTTO, B., COLABIANCHI, N., VENA, J. E., BLAIR, S. N. and HOOKER, S. P. (2016). Patterns of sedentary behavior in US middle-age and older adults: The REGARDS study. *Med. Sci. Sports Exerc.* **48** 430–438.

DIAZ, K. M., HOWARD, V. J., HUTTO, B., COLABIANCHI, N., VENA, J. E., SAFFORD, M. M., BLAIR, S. N. and HOOKER, S. P. (2017). Patterns of sedentary behavior and mortality in US middle-age and older adults: A national cohort study. *Ann. Intern. Med.* **167** 465–475.

EKELUND, U., TARP, J., STEENE-JOHANNESSEN, J., HANSEN, B. H., JEFFERIS, B., FAGERLAND, M. W., WHINCUP, P., DIAZ, K. M., HOOKER, S. P. et al. (2019). Dose-response associations between accelerometry

Key words and phrases. Adaptive dose-finding, generalized PIPE, glucose monitoring, sedentary breaks, weighted posterior gain.

measured physical activity and sedentary time and all cause mortality: Systematic review and harmonised meta-analysis. *Br. Med. J.* **366** I4570.

HSU, J. C. and BERGER, R. L. (1999). Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. *J. Amer. Statist. Assoc.* **94** 468–482.

IVANOVA, A. and WANG, K. (2006). Bivariate isotonic design for dose-finding with ordered groups. *Stat. Med.* **25** 2018–2026. MR2239229 https://doi.org/10.1002/sim.2312

LEE, S. M. and CHEUNG, Y. K. (2009). Model calibration in the continual reassessment method. *Clin. Trials* **6** 227–238.

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

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

RIVIERE, M.-K., DUBOIS, F. and ZOHAR, S. (2015). Competing designs for drug combination in phase I dose-finding clinical trials. *Stat. Med.* **34** 1–12. MR3286233 https://doi.org/10.1002/sim.6094

THALL, P. F., MILLIKAN, R. E., MUELLER, P. and LEE, S.-J. (2003). Dose-finding with two agents in Phase I oncology trials. *Biometrics* **59** 487–496. MR2004253 https://doi.org/10.1111/1541-0420.00058

WAGES, N. A., CONAWAY, M. R. and O'QUIGLEY, J. (2011). Continual reassessment method for partial ordering. *Biometrics* **67** 1555–1563. MR2872406 https://doi.org/10.1111/j.1541-0420.2011.01560.x

WHEELER, G. M., SWEETING, M. J. and MANDER, A. P. (2019). A Bayesian model-free approach to combination therapy phase I trials using censored time-to-toxicity data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 309–329. MR3902996 https://doi.org/10.1111/rssc.12323

WHEELER, G. M., SWEETING, M. J., MANDER, A. P., LEE, S. M. and CHEUNG, Y. K. K. (2017). Modelling semi-attributable toxicity in dual-agent phase I trials with non-concurrent drug administrations. *Stat. Med.* **36** 225–241. MR3582970 https://doi.org/10.1002/sim.6912

YOUNG, D. R., HIVERT, M. F., ALHASSAN, S., CAMHI, S. M., FERGUSON, J. F., KATZMARZYK, P. T., LEWIS, C. E., OWEN, N., PERRY, C. K. et al. (2016). Sedentary behavior and cardiovascular morbidity and mortality: A science advisory from the American heart association. *Circulation* **134** e262–79.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1459–1475
https://doi.org/10.1214/21-AOAS1551

BAYESIAN LOCAL FALSE DISCOVERY RATE FOR SPARSE COUNT DATA WITH APPLICATION TO THE DISCOVERY OF HOTSPOTS IN PROTEIN DOMAINS

BY IRIS IVY M. GAURAN[1,a], JUNYONG PARK[2,b], ILIA RATTSEV[3,c], THOMAS A. PETERSON[4,e], MARICEL G. KANN[3,d] AND DOHWAN PARK[5,f]

[1]*Biostatistics Group, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology,* [a]*irisivy.gauran@kaust.edu.sa*

[2]*Department of Statistics, Seoul National University,* [b]*junyongpark@snu.ac.kr*

[3]*Department of Biological Sciences, University of Maryland,* [c]*rattsev1@umbc.edu,* [d]*mkann@umbc.edu*

[4]*Institute for Computational Health Science, University of California, San Francisco,* [e]*thomas.peterson@ucsf.edu*

[5]*Department of Mathematics and Statistics, University of Maryland,* [f]*dhpark@umbc.edu*

In cancer research at the molecular level, it is critical to understand which somatic mutations play an important role in the initiation or progression of cancer. Recently, studying cancer somatic variants at the protein domain level is an important area for uncovering functionally related somatic mutations. The main issue is to find the protein domain hotspots which have significantly high frequency of mutations. Multiple testing procedures are commonly used to identify hotspots; however, when data is not large enough, existing methods produce unreliable results with failure in controlling a given type I error rate. We propose multiple testing procedures, based on Bayesian local false discovery rate, for sparse count data and apply it in the identification of clusters of somatic mutations across entire gene families using protein domain models. In multiple testing for count data, it is not clear what kind of the null distribution should be admitted. In our proposed algorithms, we implement the zero assumption in the context of Bayesian methods to identify the null distribution for count data rather than using any theoretical null distribution. Furthermore, we also address different types of modeling of alternative distributions. The proposed fully Bayesian models are efficient when the number of count data is small ($50 \leq N < 200$) while the local false discovery rate procedures, based on the empirical Bayes, is desirable for a large number of data ($N > 800$). We provide numerical studies to show that the proposed fully Bayesian methods can control a given level of false discovery rate for small number of positions while existing approaches based on nonparametric empirical Bayes fail in controlling a false discovery rate. In addition, we present real data examples of protein domain data to select hotspots in protein domain data.

REFERENCES

ANGERS, J.-F. and BISWAS, A. (2003). A Bayesian analysis of zero-inflated generalized Poisson model. *Comput. Statist. Data Anal.* **42** 37–46. MR1963008 https://doi.org/10.1016/S0167-9473(02)00154-8

ATANASOVA, V. S., RUSSELL, R. J., WEBSTER, T. G., CAO, Q., AGARWAL, P., LIM, Y. Z., KRISHNAN, S., FUENTES, I., GUTTMANN-GRUBER, C. et al. (2019). Thrombospondin-1 is a major activator of tgf-β signaling in recessive dystrophic epidermolysis bullosa fibroblasts. *J. Invest. Dermatol.* **139** 1497–1505.

BENCHARIT, S., CUI, C. B., SIDDIQUI, A., HOWARD-WILLIAMS, E. L., SONDEK, J. ZUOBI-HASONA, K. and AUKHIL, I. (2007). Structural insights into fibronectin Type III domain-mediated signaling. *J. Mol. Biol.* **367** 303–309.

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2015). Overall objective priors. *Bayesian Anal.* **10** 189–221. MR3420902 https://doi.org/10.1214/14-BA915

Key words and phrases. Bayesian local false discovery rate, sparse count data, zero-inflated generalized Poisson, protein domains.

BIAMONTI, G. and RIVA, S. (1994). New insights into the auxiliary domains of eukaryotic rna binding proteins. *FEBS Lett*. **340** 1–8.

CASSOLA, A., NOÉ, G. and FRASCH, A. C. (2010). RNA recognition motifs involved in nuclear import of RNA-binding proteins. *RNA Biology* **7** 339–344.

CHOTHIA, C. (1992). One thousand families for the molecular biologist. *Nature* **357** 543–544.

CHRISTOFORIDES, A., CARPTEN, J. D., WEISS, G. J., DEMEURE, M. J., VON HOFF, D. D. and CRAIG, D. W. (2013). Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. *BMC Genomics* **14** 302.

CIPOLLI, W. III, HANSON, T. and MCLAIN, A. C. (2016). Bayesian nonparametric multiple testing. *Comput. Statist. Data Anal.* **101** 64–79. MR3504836 https://doi.org/10.1016/j.csda.2016.02.016

DING, L., WENDL, M. C., KOBOLDT, D. C. and MARDIS, E. R. (2010). Analysis of next-generation genomic data in cancer: Accomplishments and challenges. *Hum. Mol. Genet.* **19(R2)** R188–R196.

EFRON, B. and TIBSHIRANI, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **23** 70–86.

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

GAURAN, I. I. M., PARK, J., LIM, J., PARK, D., ZYLSTRA, J., PETERSON, T., KANN, M. and SPOUGE, J. L. (2018). Empirical null estimation using zero-inflated discrete mixture distributions and its applications to protein domain data. *Biometrics* **74** 458–471. MR3825332 https://doi.org/10.1111/biom.12779

GAURAN, I. I. M., PARK, J., RATTSEV, I., PETERSON, T. A., KANN, M. G. and PARK, D. (2022). Supplement to "Bayesian local false discovery rate for sparse count data with application to the discovery of hotspots in protein domains." https://doi.org/10.1214/21-AOAS1551SUPP

GUTIÉRREZ, L., BARRIENTOS, A. F., GONZÁLEZ, J. and TAYLOR-RODRÍGUEZ, D. (2019). A Bayesian nonparametric multiple testing procedure for comparing several treatments against a control. *Bayesian Anal.* **14** 649–675. MR3959876 https://doi.org/10.1214/18-BA1122

HYNES, R. O. (2012). *Fibronectins*. Springer.

IBRAHIM, J. G., CHEN, M.-H. and GRAY, R. J. (2002). Bayesian models for gene expression with DNA microarray data. *J. Amer. Statist. Assoc.* **97** 88–99. MR1947273 https://doi.org/10.1198/016214502753479257

JOE, H. and ZHU, R. (2005). Generalized Poisson distribution: The property of mixture of Poisson and comparison with negative binomial distribution. *Biom. J.* **47** 219–229. MR2137236 https://doi.org/10.1002/bimj.200410102

JONES, S. (2004). An overview of the basic helix-loop-helix proteins. *Genome Biol.* **5** 1–6.

LARSON, D. E., HARRIS, C. C., CHEN, K., KOBOLDT, D. C., ABBOTT, T. E., DOOLING, D. J., LEY, T. J., MARDIS, E. R., WILSON, R. K. et al. (2011). SomaticSniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28** 311–317.

LI, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27** 2987–2993.

LI, R., LI, Y., KRISTIANSEN, K. and WANG, J. (2008). SOAP: Short oligonucleotide alignment program. *Bioinformatics* **24** 713–714.

MASSARI, M. E. and MURRE, C. (2000). Helix-loop-helix proteins: Regulators of transcription in eucaryotic organisms. *Mol. Cell. Biol.* **20** 429–440.

MÜLLER, P., PARMIGIANI, G., ROBERT, C. and ROUSSEAU, J. (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *J. Amer. Statist. Assoc.* **99** 990–1001. MR2109489 https://doi.org/10.1198/016214504000001646

MURRE, C., BAIN, G., VAN DIJK, M. A., ENGEL, I., FURNARI, B. A., MASSARI, M. E., MATTHEWS, J. R., QUONG, M. W., RIVERA, R. R. et al. (1994). Structure and function of helix-loop-helix proteins. *Biochimica et Biophysica Acta* (*BBA*)*-Gene Structure and Expression* **1218** 129–135.

NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.

PETERSON, T. A., PARK, D. and KANN, M. G. (2013). A protein domain-centric approach for the comparative analysis of human and yeast phenotypically relevant mutations. *BMC Genomics* **14** S5.

PETERSON, T. A., GAURAN, I. I. M., PARK, J., PARK, D. and KANN, M. G. (2017). Oncodomains: A protein domain-centric framework for analyzing rare variants in tumor samples. *PLoS Comput. Biol.* **13** e1005428. https://doi.org/10.1371/journal.pcbi.1005428

SAUNDERS, C. T., WONG, W. S. W., SWAMY, S., BECQ, J., MURRAY, L. J. and CHEETHAM, R. K. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28** 1811–1817.

SCOTT, J. G. and BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference* **136** 2144–2162. MR2235051 https://doi.org/10.1016/j.jspi.2005.08.031

SELANDER-SUNNERHAGEN, M., ULLNER, M., PERSSON, E., TELEMAN, O., STENFLO, J. and DRAKENBERG, T. (1992). How an epidermal growth factor (EGF)-like domain binds calcium. High resolution NMR structure of the calcium form of the NH2-terminal EGF-like domain in coagulation factor X. *J. Biol. Chem.* **267** 19642–19649.

SHIRAISHI, Y., SATO, Y., CHIBA, K., OKUNO, Y., NAGATA, Y., YOSHIDA, K., SHIBA, N., HAYASHI, Y., KUME, H. et al. (2013). An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.* **41** e89–e89.

YAU, C. (2013). OncoSNP-SEQ: A statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics* **29** 2482–2484.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1476–1499
https://doi.org/10.1214/21-AOAS1552

DIRICHLET-TREE MULTINOMIAL MIXTURES FOR CLUSTERING MICROBIOME COMPOSITIONS

BY JIALIANG MAO[a] AND LI MA[b]

Department of Statistical Science, Duke University, [a]*jialiang.mao@duke.com,* [b]*li.ma@duke.edu*

Studying the human microbiome has gained substantial interest in recent years, and a common task in the analysis of these data is to cluster microbiome compositions into subtypes. This subdivision of samples into subgroups serves as an intermediary step in achieving personalized diagnosis and treatment. In applying existing clustering methods to modern microbiome studies, including the American Gut Project (AGP) data, we found that this seemingly standard task, however, is very challenging in the microbiome composition context, due to several key features of such data. Standard distance-based clustering algorithms generally do not produce reliable results, as they do not take into account the heterogeneity of the cross-sample variability among the bacterial taxa, while existing model-based approaches do not allow sufficient flexibility for the identification of complex within-cluster variation from cross-cluster variation. Direct applications of such methods generally lead to overly dispersed clusters in the AGP data, and such a phenomenon is common for other microbiome data. To overcome these challenges, we introduce Dirichlet-tree multinomial mixtures (DTMM) as a Bayesian generative model for clustering amplicon sequencing data in microbiome studies. DTMM models the microbiome population with a mixture of Dirichlet-tree kernels that utilizes the phylogenetic tree to offer a more flexible covariance structure in characterizing within-cluster variation, and it provides a means for identifying a subset of signature taxa that distinguish the clusters. We perform extensive simulation studies to evaluate the performance of DTMM, compare it to state-of-the-art model-based and distance-based clustering methods in the microbiome context and carry out a validation study on a publicly available longitudinal data set to confirm the biological relevance of the clusters. Finally, we report a case study on the fecal data from the AGP to identify compositional clusters among individuals with inflammatory bowel disease and diabetes. Among our most interesting findings is that enterotypes (i.e., gut microbiome clusters) are not always defined by the most dominant taxa, as previous analyses had assumed, but can involve a number of less abundant taxa which cannot be identified with existing distance-based and method-based approaches.

REFERENCES

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

ANDERSON, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26** 32–46.

ARUMUGAM, M., RAES, J., PELLETIER, E., LE PASLIER, D., YAMADA, T., MENDE, D. R., FERNANDES, G. R., TAP, J., BRULS, T. et al. (2011). Enterotypes of the human gut microbiome. *Nature* **473** 174.

CALLAHAN, B. J., MCMURDIE, P. J. and HOLMES, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11** 2639–2643. https://doi.org/10.1038/ismej.2017.119

CALLAHAN, B. J., MCMURDIE, P. J., ROSEN, M. J., HAN, A. W., JOHNSON, A. J. A. and HOLMES, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13** 581.

Key words and phrases. Bayesian hierarchical models, compositional data, latent variable models, probabilistic learning.

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

COSTEA, P. I., HILDEBRAND, F., ARUMUGAM, M., BÄCKHED, F., BLASER, M. J., BUSHMAN, F. D., DE VOS, W. M., EHRLICH, S. D., FRASER, C. M. et al. (2018). Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* **3** 8–16.

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

DENNIS, S. Y. III (1991). On the hyper-Dirichlet type 1 and hyper-Liouville distributions. *Comm. Statist. Theory Methods* **20** 4069–4081. MR1158563 https://doi.org/10.1080/03610929108830757

DETHLEFSEN, L. and RELMAN, D. A. (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. USA* **108** 4554–4561.

HOLMES, I., HARRIS, K. and QUINCE, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS ONE* **7** e30126. https://doi.org/10.1371/journal.pone.0030126

ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. MR1952729 https://doi.org/10.1198/016214501750332758

JACCARD, P. (1912). The distribution of the flora in the Alpine zone. 1. *New Phytol.* **11** 37–50.

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

KAUFMAN, L. and ROUSSEEUW, P. J. (2009). *Finding Groups in Data*: *An Introduction to Cluster Analysis* **344**. Wiley, New York.

KNIGHTS, D., KUCZYNSKI, J., CHARLSON, E. S., ZANEVELD, J., MOZER, M. C., COLLMAN, R. G., BUSHMAN, F. D., KNIGHT, R. and KELLEY, S. T. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **8** 761.

KOREN, O., KNIGHTS, D., GONZALEZ, A., WALDRON, L., SEGATA, N., KNIGHT, R., HUTTENHOWER, C. and LEY, R. E. (2013). A guide to enterotypes across the human body: Meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* **9** e1002863. https://doi.org/10.1371/journal.pcbi.1002863

KOSTIC, A. D., XAVIER, R. J. and GEVERS, D. (2014). The microbiome in inflammatory bowel disease: Current status and the future ahead. *Gastroenterology* **146** 1489–1499.

KUNTZ, T. M. and GILBERT, J. A. (2017). Introducing the microbiome into precision medicine. *Trends Pharmacol. Sci.* **38** 81–91. https://doi.org/10.1016/j.tips.2016.10.001

LA ROSA, P. S., BROOKS, J. P., DEYCH, E., BOONE, E. L., EDWARDS, D. J., WANG, Q., SODERGREN, E., WEINSTOCK, G. and SHANNON, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE* **7** e52078.

LLOYD, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28** 129–137. MR0651807 https://doi.org/10.1109/TIT.1982.1056489

LOZUPONE, C. and KNIGHT, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71** 8228–8235.

MA, L. and SORIANO, J. (2018). Analysis of distributional variation through graphical multi-scale beta-binomial models. *J. Comput. Graph. Statist.* **27** 529–541. MR3863755 https://doi.org/10.1080/10618600.2017.1402774

MAO, J. and MA, L. (2022). Supplement to "Dirichlet-tree multinomial mixtures for clustering microbiome compositions." https://doi.org/10.1214/21-AOAS1552SUPPA, https://doi.org/10.1214/21-AOAS1552SUPPB

MAO, J., CHEN, Y. and MA, L. (2020). Bayesian graphical compositional regression for microbiome data. *J. Amer. Statist. Assoc.* **115** 610–624. MR4107661 https://doi.org/10.1080/01621459.2019.1647212

MCDONALD, D., BIRMINGHAM, A. and KNIGHT, R. (2015). Context and the human microbiome. *Microbiome* **3** 52. https://doi.org/10.1186/s40168-015-0117-2

MCDONALD, D., HYDE, E., DEBELIUS, J. W., MORTON, J. T., GONZALEZ, A., ACKERMANN, G., AKSENOV, A. A., BEHSAZ, B., BRENNAN, C. et al. (2018). American gut: An open platform for citizen science microbiome research. *MSystems* **3** e00031–18.

MILLER, J. W. and HARRISON, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems* 199–206.

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

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

NG, A. Y., JORDAN, M. I. and WEISS, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* 849–856.

QIN, J., LI, Y., CAI, Z., LI, S., ZHU, J., ZHANG, F., LIANG, S., ZHANG, W., GUAN, Y. et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490** 55.

QUINCE, C., LUNDIN, E. E., ANDREASSON, A. N., GRECO, D., RAFTER, J., TALLEY, N. J., AGREUS, L., ANDERSSON, A. F., ENGSTRAND, L. et al. (2013). The impact of Crohn's disease genes on healthy human gut microbiota: A pilot study. *Gut* **62** 952–954.

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

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

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

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

WANG, T. and ZHAO, H. (2017). A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics* **73** 792–801. MR3713113 https://doi.org/10.1111/biom.12654

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.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1500–1523
https://doi.org/10.1214/21-AOAS1553

SEMIPARAMETRIC POINT PROCESS MODELING OF BLINKING ARTIFACTS IN PALM

BY LOUIS G. JENSEN[1,a], DAVID J. WILLIAMSON[2,c] AND UTE HAHN[1,b]

[1]*Department of Mathematics, Aarhus University,* [a]*Louis.Gammelgaard@gmail.com,* [b]*ute@math.au.dk*
[2]*Randall Division for Cell and Molecular Biophysics, King's College London,* [c]*david.williamson@kcl.ac.uk*

Photoactivated localization microscopy (PALM) is a powerful imaging technique for characterization of protein organization in biological cells. Due to the stochastic blinking of fluorescent probes and camera discretization effects, each protein gives rise to a cluster of artificial observations. These blinking artifacts are an obstacle for quantitative analysis of PALM data, and tools for their correction are in high demand. We develop the independent blinking cluster point process (IBCpp) family of models, which is suited for modeling of data from single-molecule localization microscopy modalities, and we present results on the mark correlation function. We then construct the PALM-IBCpp, a semiparametric IBCpp tailored for PALM data, and we describe a procedure for estimation of parameters which can be used without parametric assumptions on the spatial organization of proteins. Our model is validated on nuclear pore complex reference data, where the ground truth was accurately recovered, and we demonstrate how the estimated blinking parameters can be used to perform a blinking corrected test for protein clustering in a cell expressing the adaptor protein LAT. Finally, we consider simulations with varying degrees of blinking and protein clustering to shed light on the expected performance in a range of realistic settings.

REFERENCES

ANDERSEN, I. T., HAHN, U., ARNSPANG, E. C., NEJSUM, L. N. and JENSEN, E. B. V. (2018). Double Cox cluster processes—With applications to photoactivated localization microscopy. *Spat. Stat.* **27** 58–73. MR3868199 https://doi.org/10.1016/j.spasta.2018.04.009

ANNIBALE, P., VANNI, S., SCARSELLI, M., ROTHLISBERGER, U. and RADENOVIC, A. (2011a). Quantitative photo activated localization microscopy: Unraveling the effects of photoblinking. *PLoS ONE* **6** e22678. https://doi.org/10.1371/journal.pone.0022678

ANNIBALE, P., VANNI, S., SCARSELLI, M., ROTHLISBERGER, U. and RADENOVIC, A. (2011b). Identification of clustering artifacts in photoactivated localization microscopy. *Nat. Methods* **8** 527–528. https://doi.org/10.1038/nmeth.1627

BETZIG, E., PATTERSON, G. H., SOUGRAT, R., LINDWASSER, O. W., OLENYCH, S., BONIFACINO, J. S., DAVIDSON, M. W., LIPPINCOTT-SCHWARTZ, J. and HESS, H. F. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313** 1642–1645.

COLTHARP, C., KESSLER, R. P. and XIAO, J. (2012). Accurate construction of photoactivated localization microscopy (PALM) images for quantitative measurements. *PLoS ONE* **7** e51725. https://doi.org/10.1371/journal.pone.0051725

DALEY, D. J. and VERE-JONES, D. (2007). *An Introduction to the Theory of Point Processes. Vol. II: General Theory and Structure*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR2371524 https://doi.org/10.1007/978-0-387-49835-5

DESCHOUT, H., ZANACCHI, F. C., MLODZIANOSKI, M., DIASPRO, A., BEWERSDORF, J., HESS, S. T. and BRAECKMANS, K. (2014). Precisely and accurately localizing single emitters in fluorescence microscopy. *Nat. Methods* **11** 253.

FRICKE, F., BEAUDOUIN, J., EILS, R. and HEILEMANN, M. (2015). One, two or three? Probing the stoichiometry of membrane proteins by single-molecule localization microscopy. *Sci. Rep.* **5** 14072.

Key words and phrases. Photoactivated localization microscopy, multiple blinking, spatiotemporal point patterns, mark correlation function, moment-based estimation, second-order characteristics.

GRIFFIÉ, J., PHAM, T. A., SIEBEN, C., LANG, R., CEVHER, V., HOLDEN, S., UNSER, M., MANLEY, S. and SAGE, D. (2020). Virtual-SMLM, a virtual environment for real-time interactive SMLM acquisition. https://doi.org/10.1101/2020.03.05.967893

HUANG, B., BATES, M. and ZHUANG, X. (2009). Super-resolution fluorescence microscopy. *Annu. Rev. Biochem.* **78** 993–1016. https://doi.org/10.1146/annurev.biochem.77.061906.092014

HUMMER, G., FRICKE, F. and HEILEMANN, M. (2016). Model-independent counting of molecules in single-molecule localization microscopy. *Mol. Biol. Cell* **27** 3637–3644. https://doi.org/10.1091/mbc.E16-07-0525

JENSEN, L. G., WILLIAMSON, D. J. and HAHN, U. (2022). Supplement to "Semiparametric point process modeling of blinking artifacts in PALM." https://doi.org/10.1214/21-AOAS1553SUPPA, https://doi.org/10.1214/21-AOAS1553SUPPB

KARATHANASIS, C., FRICKE, F., HUMMER, G. and HEILEMANN, M. (2017). Molecule counts in localization microscopy with organic fluorophores. *ChemPhysChem* **18** 942–948. https://doi.org/10.1002/cphc.201601425

KHATER, I. M., NABI, I. R. and HAMARNEH, G. (2020). A review of super-resolution single-molecule localization microscopy cluster analysis and quantification methods. *Patterns* **1** 100038. https://doi.org/10.1016/j.patter.2020.100038

LEE, S. H., SHIN, J. Y., LEE, A. and BUSTAMANTE, C. (2012). Counting single photoactivatable fluorescent molecules by photoactivated localization microscopy (PALM). *Proc. Natl. Acad. Sci. USA* **109** 17436–17441. https://doi.org/10.1073/pnas.1215175109

LIN, Y., LONG, J. J., HUANG, F., DUIM, W. C., KIRSCHBAUM, S., ZHANG, Y., SCHROEDER, L. K., REBANE, A. A., VELASCO, M. G. M. et al. (2015). Quantifying and optimizing single-molecule switching nanoscopy at high speeds. *PLoS ONE* **10** e0128135.

MYLLYMÄKI, M., MRKVIČKA, T., GRABARNIK, P., SEIJO, H. and HAHN, U. (2017). Global envelope tests for spatial processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 381–404. MR3611751 https://doi.org/10.1111/rssb.12172

OBER, R. J., TAHMASBI, A., RAM, S., LIN, Z. and WARD, E. S. (2015). Quantitative aspects of single-molecule microscopy: Information-theoretic analysis of single-molecule data. *IEEE Signal Process. Mag.* **32** 58–69. https://doi.org/10.1109/msp.2014.2353664

OVESNÝ, M., KŘÍŽEK, P., BORKOVEC, J., ŠVINDRYCH, Z. and HAGEN, G. M. (2014). ThunderSTORM: A comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics* **30** 2389–2390.

PATEL, L., GUSTAFSSON, N., LIN, Y., OBER, R., HENRIQUES, R. and COHEN, E. (2019). A hidden Markov model approach to characterizing the photo-switching behavior of fluorophores. *Ann. Appl. Stat.* **13** 1397–1429. MR4019144 https://doi.org/10.1214/19-AOAS1240

RIES, J. (2020). SMAP: A modular super-resolution microscopy analysis platform for SMLM data. *Nat. Methods* **17** 870–872. https://doi.org/10.1038/s41592-020-0938-1

RIPLEY, B. D. (1976). The second-order analysis of stationary point processes. *J. Appl. Probab.* **13** 255–266. MR0402918 https://doi.org/10.1017/s0021900200094328

ROLLINS, G. C., SHIN, J. Y., BUSTAMANTE, C. and PRESSÉ, S. (2015). Stochastic approach to the molecular counting problem in superresolution microscopy. *Proc. Natl. Acad. Sci. USA* **112** E110–E118.

ROSSBOTH, B., ARNOLD, A. M., TA, H., PLATZER, R., KELLNER, F., HUPPA, J. B., BRAMESHUBER, M., BAUMGART, F. and SCHÜTZ, G. J. (2018). TCRs are randomly distributed on the plasma membrane of resting antigen-experienced T cells. *Nat. Immunol.* **19** 821–827. https://doi.org/10.1038/s41590-018-0162-7

RUST, M. J., BATES, M. and ZHUANG, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3** 793–796. https://doi.org/10.1038/nmeth929

SCHNITZBAUER, J., STRAUSS, M. T., SCHLICHTHAERLE, T., SCHUEDER, F. and JUNGMANN, R. (2017). Super-resolution microscopy with DNA-PAINT. *Nat. Protoc.* **12** 1198–1228. https://doi.org/10.1038/nprot.2017.024

SENGUPTA, P., JOVANOVIC-TALISMAN, T., SKOKO, D., RENZ, M., VEATCH, S. L. and LIPPINCOTT-SCHWARTZ, J. (2011). Probing protein heterogeneity in the plasma membrane using PALM and pair correlation analysis. *Nat. Methods* **8** 969–975. https://doi.org/10.1038/nmeth.1704

SHIVANANDAN, A., DESCHOUT, H., SCARSELLI, M. and RADENOVIC, A. (2014). Challenges in quantitative single molecule localization microscopy. *FEBS Lett.* **588** 3595–3602.

SHTENGEL, G., GALBRAITH, J. A., GALBRAITH, C. G., LIPPINCOTT-SCHWARTZ, J., GILLETTE, J. M., MANLEY, S., SOUGRAT, R., WATERMAN, C. M., KANCHANAWONG, P. et al. (2009). Interferometric fluorescent super-resolution microscopy resolves 3D cellular ultrastructure. *Proc. Natl. Acad. Sci. USA* **106** 3125–3130. https://doi.org/10.1073/pnas.0813131106

SMALL, A. and STAHLHEBER, S. (2014). Fluorophore localization algorithms for super-resolution microscopy. *Nat. Methods* **11** 267–279. https://doi.org/10.1038/nmeth.2844

STAUDT, T., ASPELMEIER, T., LAITENBERGER, O., GEISLER, C., EGNER, A. and MUNK, A. (2020). Statistical molecule counting in super-resolution fluorescence microscopy: Towards quantitative nanoscopy. *Statist. Sci.* **35** 92–111. MR4071360 https://doi.org/10.1214/19-STS753

STOYAN, D. (1984). On correlations of marked point processes. *Math. Nachr.* **116** 197–207. MR0762601 https://doi.org/10.1002/mana.19841160115

THEVATHASAN, J. V., KAHNWALD, M., CIEŚLIŃSKI, K., HOESS, P., PENETI, S. K., REITBERGER, M., HEID, D., KASUBA, K. C., HOERNER, S. J. et al. (2019a). Nuclear pores as versatile reference standards for quantitative superresolution microscopy. *Nat. Methods* **16** 1045–1053. https://doi.org/10.1038/s41592-019-0574-9

THEVATHASAN, J. V., KAHNWALD, M., CIEŚLIŃSKI, K., HOESS, P., PENETI, S. K., REITBERGER, M., HEID, D., KASUBA, K. C., HOERNER, S. J. et al. (2019b). Nuclear pores as versatile reference standards for quantitative superresolution microscopy. Available at https://www.ebi.ac.uk/biostudies/BioImages/studies/S-BIAD8.

VEATCH, S. L., MACHTA, B. B., SHELBY, S. A., CHIANG, E. N., HOLOWKA, D. A. and BAIRD, B. A. (2012). Correlation functions quantify super-resolution images and estimate apparent clustering due to over-counting. *PLoS ONE* **7** e31457. https://doi.org/10.1371/journal.pone.0031457

WILLIAMSON, D. J., OWEN, D. M., ROSSY, J., MAGENAU, A., WEHRMANN, M., GOODING, J. J. and GAUS, K. (2011). Pre-existing clusters of the adaptor Lat do not participate in early T cell signaling events. *Nat. Immunol.* **12** 655–662. https://doi.org/10.1038/ni.2049

YAMANAKA, M., SMITH, N. I. and FUJITA, K. (2014). Introduction to super-resolution microscopy. *Microscopy* **63** 177–192.

ZHANG, B., ZERUBIA, J. and OLIVO-MARIN, J.-C. (2007). Gaussian approximations of fluorescence microscope point-spread function models. *Appl. Opt.* **46** 1819–1829.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1524–1549
https://doi.org/10.1214/21-AOAS1555

IMPROVED INFERENCE ON RISK MEASURES FOR UNIVARIATE EXTREMES

BY LÉO R. BELZILE[1,a] AND ANTHONY C. DAVISON[2,b]

[1]*Department of Decision sciences, HEC Montréal,* [a]*leo.belzile@hec.ca*
[2]*Institute of Mathematics, École polytechnique fédérale de Lausanne,* [b]*anthony.davison@epfl.ch*

We discuss the use of likelihood asymptotics for inference on risk measures in univariate extreme value problems, focusing on estimation of high quantiles and similar summaries of risk for uncertainty quantification. We study whether higher-order approximation, based on the tangent exponential model, can provide improved inferences. We conclude that inference based on maxima is generally robust to mild model misspecification and that profile likelihood-based confidence intervals will often be adequate, whereas inferences based on threshold exceedances can be badly biased but may be improved by higher-order methods, at least for moderate sample sizes. We use the methods to shed light on catastrophic rainfall in Venezuela, flooding in Venice, and the lifetimes of Italian semisupercentenarians.

REFERENCES

BARBI, E., LAGONA, F., MARSILI, M., VAUPEL, J. W. and WACHTER, K. W. (2018). The plateau of human mortality: Demography of longevity pioneers. *Science* **360** 1459–1461. MR3792342 https://doi.org/10.1126/science.aat3119

BARLOW, A. M., SHERLOCK, C. and TAWN, J. (2020). Inference for extreme values under threshold-based stopping rules. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 765–789. MR4133146

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1994). *Inference and Asymptotics*. *Monographs on Statistics and Applied Probability* **52**. CRC Press, London. MR1317097 https://doi.org/10.1007/978-1-4899-3210-5

BELZILE, L. R. (2019). Contributions to likelihood-based modelling of extreme values. Ph.D. thesis. EPFL, Lausanne.

BELZILE, L. R and DAVISON, A. C (2022). Supplement to "Improved inference on risk measures for univariate extremes." https://doi.org/10.1214/21-AOAS1555SUPPA, https://doi.org/10.1214/21-AOAS1555SUPPB

BELZILE, L. R., DAVISON, A. C., ROOTZÉN, H. and ZHOLUD, D. (2021). Human mortality at extreme age. *R. Soc. Open Sci.* **8** 202097. https://doi.org/10.1098/rsos.202097

BELZILE, L. R., DAVISON, A. C., GAMPE, J., ROOTZÉN, H. and ZHOLUD, D. (2022). Is there a cap on longevity? A statistical review. *Annu. Rev. Stat. Appl.* **9**, in press. https://doi.org/10.1146/annurev-statistics-040120-025426

BRAZZALE, A. R., DAVISON, A. C. and REID, N. (2007). *Applied Asymptotics*: *Case Studies in Small-Sample Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **23**. Cambridge Univ. Press, Cambridge. MR2342742 https://doi.org/10.1017/CBO9780511611131

BÜCHER, A. and SEGERS, J. (2017). On the maximum likelihood estimator for the generalized extreme-value distribution. *Extremes* **20** 839–872. MR3737387 https://doi.org/10.1007/s10687-017-0292-6

BUITENDAG, S., BEIRLANT, J. and DE WET, T. (2020). Confidence intervals for extreme Pareto-type quantiles. *Scand. J. Stat.* **47** 36–55. MR4075228 https://doi.org/10.1111/sjos.12396

COLES, S. and PERICCHI, L. (2003). Anticipating catastrophes through extreme value modelling. *J. Roy. Statist. Soc. Ser. C* **52** 405–416. MR2012566 https://doi.org/10.1111/1467-9876.00413

COLES, S., PERICCHI, L. R. and SISSON, S. (2003). A fully probabilistic approach to extreme rainfall modeling. *J. Hydrol.* **273** 35–50.

COX, D. R., ISHAM, V. S. and NORTHROP, P. J. (2002). Floods: Some probabilistic and statistical approaches. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **360** 1389–1408.

COX, D. R. and SNELL, E. J. (1968). A general definition of residuals (with discussion). *J. Roy. Statist. Soc. Ser. B* **30** 248–275. MR0237052

Key words and phrases. Extreme value distribution, generalized Pareto distribution, higher-order asymptotic inference, Poisson process, tangent exponential model, profile likelihood.

DAVIS, R. A. and MIKOSCH, T. (2009). The extremogram: A correlogram for extreme events. *Bernoulli* **15** 977–1009. MR2597580 https://doi.org/10.3150/09-BEJ213

DAVISON, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73** 323–332. MR0855892 https://doi.org/10.1093/biomet/73.2.323

DAVISON, A. C. (2003). *Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **11**. Cambridge Univ. Press, Cambridge. MR1998913 https://doi.org/10.1017/CBO9780511815850

DAVISON, A. C., FRASER, D. A. S. and REID, N. (2006). Improved likelihood inference for discrete data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 495–508. MR2278337 https://doi.org/10.1111/j.1467-9868.2006.00548.x

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

DAVISON, A. C. and REID, N. (2022). The tangent exponential model. In *Handbook of Bayesian, Fiducial and Frequentist Inference* (J. O. Berger, X. L. Meng, N. Reid and M. Xie, eds.) CRC Press/CRC, Boca Raton, FL.

DE CARVALHO, M. and DAVISON, A. C. (2014). Spectral density ratio models for multivariate extremes. *J. Amer. Statist. Assoc.* **109** 764–776. MR3223748 https://doi.org/10.1080/01621459.2013.872651

DE HAAN, L. and FERREIRA, A. (2006). *Extreme Value Theory: An Introduction*. *Springer Series in Operations Research and Financial Engineering*. Springer, New York. MR2234156 https://doi.org/10.1007/0-387-34471-3

DOMBRY, C. and FERREIRA, A. (2019). Maximum likelihood estimators based on the block maxima method. *Bernoulli* **25** 1690–1723. MR3961227 https://doi.org/10.3150/18-BEJ1032

EINMAHL, J. J., EINMAHL, J. H. J. and DE HAAN, L. (2019). Limits to human life span through extreme value theory. *J. Amer. Statist. Assoc.* **114** 1075–1080. MR4011759 https://doi.org/10.1080/01621459.2018.1537912

EINMAHL, J. H. J. and SEGERS, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *Ann. Statist.* **37** 2953–2989. MR2541452 https://doi.org/10.1214/08-AOS677

EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (1997). *Modelling Extremal Events—for Insurance and Finance*. *Applications of Mathematics* **33**. Springer, Berlin. MR1458613 https://doi.org/10.1007/978-3-642-33483-2

FASIOLO, M., WOOD, S. N., ZAFFRAN, M., NEDELLEC, R. and GOUDE, Y. (2021). Fast calibrated additive quantile regression. *J. Amer. Statist. Assoc.* **116** 1402–1412. MR4309281 https://doi.org/10.1080/01621459.2020.1725521

FIGUEIREDO, F., GOMES, M. I., HENRIQUES-RODRIGUES, L. and MIRANDA, M. C. (2012). A computational study of a quasi-PORT methodology for VaR based on second-order reduced-bias estimation. *J. Stat. Comput. Simul.* **82** 587–602. MR2908951 https://doi.org/10.1080/00949655.2010.547196

FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80** 27–38. MR1225212 https://doi.org/10.1093/biomet/80.1.27

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

FRASER, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? *Statist. Sci.* **26** 299–316. MR2918001 https://doi.org/10.1214/11-STS352

FRASER, D. A. S., REID, N. and WU, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86** 249–264. MR1705367 https://doi.org/10.1093/biomet/86.2.249

FRASER, D. A. S., WONG, A. and WU, J. (1999). Regression analysis, nonlinear or nonnormal: Simple and accurate p values from likelihood analysis. *J. Amer. Statist. Assoc.* **94** 1286–1295. MR1731490 https://doi.org/10.2307/2669942

FRASER, D. A. S., BÉDARD, M., WONG, A., LIN, W. and FRASER, A. M. (2016). Bayes, reproducibility and the quest for truth. *Statist. Sci.* **31** 578–590. MR3598740 https://doi.org/10.1214/16-STS573

GILES, D. E., FENG, H. and GODWIN, R. T. (2016). Bias-corrected maximum likelihood estimation of the parameters of the generalized Pareto distribution. *Comm. Statist. Theory Methods* **45** 2465–2483. MR3480663 https://doi.org/10.1080/03610926.2014.887104

GNEDENKO, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. of Math.* (2) **44** 423–453. MR0008655 https://doi.org/10.2307/1968974

GOMES, M. I. and PESTANA, D. (2007). A sturdy reduced-bias extreme quantile (VaR) estimator. *J. Amer. Statist. Assoc.* **102** 280–292. MR2345543 https://doi.org/10.1198/016214506000000799

HANAYAMA, N. and SIBUYA, M. (2016). Estimating the upper limit of lifetime probability distribution, based on data of Japanese centenarians. *J. Gerontol., Ser. A* **71** 1014–1021.

HILL, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3** 1163–1174. MR0378204

HOSKING, J. R. M. and WALLIS, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics* **29** 339–349. MR0906643 https://doi.org/10.2307/1269343

KENNE PAGUI, E. C., SALVAN, A. and SARTORI, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika* **104** 923–938. MR3737312 https://doi.org/10.1093/biomet/asx046

LEE, S. M. S. and YOUNG, G. A. (2005). Parametric bootstrapping with nuisance parameters. *Statist. Probab. Lett.* **71** 143–153. MR2126770 https://doi.org/10.1016/j.spl.2004.10.026

MHALLA, L., DE CARVALHO, M. and CHAVEZ-DEMOULIN, V. (2019). Regression-type models for extremal dependence. *Scand. J. Stat.* **46** 1141–1167. MR4033807 https://doi.org/10.1111/sjos.12388

PICKANDS, J. III (1986). The continuous and differentiable domains of attraction of the extreme value distributions. *Ann. Probab.* **14** 996–1004. MR0841599

PIRAZZOLI, P. A. (1982). Maree estreme a Venezia (periodo 1872–1981). *Acqua Aria* **10** 1023–1039.

PIRES, J. F., CYSNEIROS, A. H. M. A. and CRIBARI-NETO, F. (2018). Improved inference for the generalized Pareto distribution. *Braz. J. Probab. Stat.* **32** 69–85. MR3770864 https://doi.org/10.1214/16-BJPS332

R CORE TEAM (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

ROODMAN, D. (2018). Bias and size corrections in extreme value modeling. *Comm. Statist. Theory Methods* **47** 3377–3391. MR3803408 https://doi.org/10.1080/03610926.2017.1353630

ROOTZÉN, H. and ZHOLUD, D. (2017). Human life is unlimited—but short (with discussion). *Extremes* **20** 713–728. MR3737382 https://doi.org/10.1007/s10687-017-0305-5

SEVERINI, T. A. (2000). *Likelihood Methods in Statistics. Oxford Statistical Science Series* **22**. Oxford Univ. Press, Oxford. MR1854870

SKOVGAARD, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2** 145–165. MR1410135 https://doi.org/10.2307/3318548

SMITH, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72** 67–90. MR0790201 https://doi.org/10.1093/biomet/72.1.67

SMITH, R. L. (1986). Extreme value theory based on the r largest annual events. *J. Hydrol.* **86** 27–43.

SMITH, R. L. (1987). Approximations in extreme value theory. Technical Report 205, Center for Stochastic Processes, University of North Carolina Chapel Hill.

THE SAGE DEVELOPERS (2021). SageMath, the Sage Mathematics Software System (Version 9.3).

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. MR0830567

TIERNEY, L., KASS, R. E. and KADANE, J. B. (1989). Approximate marginal densities of nonlinear functions. *Biometrika* **76** 425–433 (correction: **78**, 233–234). MR1040637 https://doi.org/10.1093/biomet/76.3.425

WADSWORTH, J. L. (2016). Exploiting structure of maximum likelihood estimators for extreme value threshold selection. *Technometrics* **58** 116–126. MR3463162 https://doi.org/10.1080/00401706.2014.998345

WANG, H. and TSAI, C.-L. (2009). Tail index regression. *J. Amer. Statist. Assoc.* **104** 1233–1240. MR2750246 https://doi.org/10.1198/jasa.2009.tm08458

WEISSMAN, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Amer. Statist. Assoc.* **73** 812–815. MR0521329

WIECZOREK, G. F., LARSEN, M. C., EATON, L. S., MORGAN, B. A. and BLAIR, J. L. (2001). Debris-flow and flooding hazards associated with the December 1999 storm in coastal Venezuela and strategies for mitigation Technical Report No. 01-0144 U.S. Geological Survey.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1550–1562
https://doi.org/10.1214/21-AOAS1556

A BAYESIAN HIERARCHICAL MODEL FOR COMBINING MULTIPLE DATA SOURCES IN POPULATION SIZE ESTIMATION

BY JACOB PARSONS[1,a], XIAOYUE NIU[2,b] AND LE BAO[2,c]

[1]*GlaxoSmithKline,* [a]*contact@jacobleeparsons.com*
[2]*Department of Statistics, Pennsylvania State University,* [b]*xiaoyue@psu.edu,* [c]*lebao@psu.edu*

To combat the HIV/AIDS pandemic effectively, targeted interventions among certain key populations play a critical role. Examples of such key populations include sex workers, people who inject drugs, and men who have sex with men. While having accurate estimates for the size of these key populations is important, any attempt to directly contact or count members of these populations is difficult. As a result, indirect methods are used to produce size estimates. Multiple approaches for estimating the size of such populations have been suggested but often give conflicting results. It is, therefore, necessary to have a principled way to combine and reconcile these estimates. To this end, we present a Bayesian hierarchical model for estimating the size of key populations that combines multiple estimates from different sources of information. The proposed model makes use of multiple years of data and explicitly models the systematic error in the data sources used. We use the model to estimate the size of people who inject drugs in Ukraine. We evaluate the appropriateness of the model and compare the contribution of each data source to the final estimates.

REFERENCES

ABDUL-QUADER, A. S., BAUGHMAN, A. L. and HLADIK, W. (2014). Estimating the size of key populations: Current status and future possibilities. *Curr. Opin. HIV AIDS* **9** 107–114.

BAO, L., RAFTERY, A. E. and REDDY, A. (2015). Estimating the sizes of populations at risk of HIV infection from multiple data sources using a Bayesian hierarchical model. *Stat. Interface* **8** 125–136. MR3322160 https://doi.org/10.4310/SII.2015.v8.n2.a1

BERLEVA, G. and SAZONOVA, Y. (2017). *Analytical Report Based on Sociological Study Results*: *Estimation of the Size of Populations Most-at-Risk for HIV Infection in Ukraine in* 2017. Alliance of Public Health.

BERLEVA, G., DUMCHEV, K., KOBYSHCHA, Y. V., PANIOTTO, V. I., PETRENKO, T. V., SALIUK, T. O. and SHVAB, I. A. (2010). *Analytical Report Based on Sociological Study Results*: *Estimation of the Size of Populations Most-at-Risk for Hiv Infection in Ukraine in* 2009. International HIV/AIDS Alliance in Ukraine, Kyiv.

BERLEVA, G., DUMCHEV, K., KASIANCHUK, M., NIKOLKO, M., SALIUK, T., SHAVB, I. and YAREMENKO, O. (2012). *Estimation of the Size of Populations Most-at-Risk for HIV Infection in Ukraine as of* 2012. International HIV/AIDS Alliance in Ukraine, Kyiv.

BERNARD, H. R., JOHNSEN, E. C., KILLWORTH, P. D. and ROBINSON, S. (1989). Estimating the size of an average personal network and of an event subpopulation. In *The Small World* 159–175. Ablex, Norwood.

BROWN, T. and PEERAPATANAPOKIN, W. (2004). The Asian epidemic model: A process model for exploring HIV policy and programme alternatives in Asia. *Sex. Transm. Infect.* **80** i19–i24.

FEEHAN, D. M. and SALGANIK, M. J. (2016). Generalizing the network scale-up method: A new estimator for the size of hidden populations. *Sociol. Method.* **46** 153–186.

FEEHAN, D. M., UMUBYEYI, A., MAHY, M., HLADIK, W. and SALGANIK, M. J. (2016). Quantity versus quality: A survey experiment to improve the network scale-up method. *Am. J. Epidemiol.* **183** 747–757.

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

GHYS, P. D., BROWN, T., GRASSLY, N. C., GARNETT, G., STANECKI, K. A., STOVER, J. and WALKER, N. (2004). The UNAIDS estimation and projection package: A software package to estimate and project national HIV epidemics. *Sex. Transm. Infect.* **80** i5–i9.

Key words and phrases. Multiplier method, network scale-up, HIV/AIDS epidemic, people who inject drugs, key affected population.

JOHNSEN, E. C., BERNARD, H. R., KILLWORTH, P. D., SHELLEY, G. A. and MCCARTY, C. (1995). A social network approach to corroborating the number of aids/hiv+ victims in the us. *Soc. Netw.* **17** 167–187.

JOHNSTON, DIMITRI, P., FISHER, R. H., ALI, M., CHOMNAD, M. and WILLI, M. (2013). Incorporating the service multiplier method in respondent-driven sampling surveys to estimate the size of hidden and hard-to-reach populations: Case studies from around the world. *Sex. Transm. Infect.* **40** 304–310.

KILLWORTH, P. D., MCCARTY, C., BERNARD, H. R., SHELLEY, G. A. and JOHNSEN, E. C. (1998). Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach. *Eval. Rev.* **22** 289–308.

LAGA, I., BAO, L. and NIU, X. (2021). Thirty years of the network scale-up method. *J. Amer. Statist. Assoc.* **116** 1548–1559. MR4309292 https://doi.org/10.1080/01621459.2021.1935267

MALTIEL, R., RAFTERY, A. and MCCORMICK, T. H. (2013). Estimating population size using the network scale up method. *Ann. Appl. Stat.* **9** 1247–1277.

MCCARTY, C., KILLWORTH, P. D., BERNARD, H. R., JOHNSEN, E. C. and SHELLEY, G. A. (2001). Comparing two methods for estimating network size. *Hum. Org.* **60** 28–39.

MCCORMICK, T. H., SALGANIK, M. J. and ZHENG, T. (2010). How many people do you know? Efficiently estimating personal network size. *J. Amer. Statist. Assoc.* **105** 59–70. MR2757192 https://doi.org/10.1198/jasa.2009.ap08518

OKAL, J., GEIBEL, S., MURAGURI, N., MUSYOKI, H., TUN, W., BROZ, D., KURIA, D., KIM, A., OLUOCH, T. et al. (2013). Estimates of the size of key populations at risk for HIV infection: Men who have sex with men, female sex workers and injecting drug users in Nairobi, Kenya. *Sex. Transm. Infect.* **89** 366–371.

PANIOTTO, V., PETRENKO, T., KUPRIYANOV, O. and PAKHOK, O. (2009). *Estimating the Size of Populations with High Risk for HIV Using the Network Scale-up Method*. Kiev International Institute of Sociology, Ukraine.

PARSONS, J., NIU, X. and BAO, L. (2022). Supplement to "A Bayesian hierarchical model for combining multiple data sources in population size estimation." https://doi.org/10.1214/21-AOAS1556SUPPA, https://doi.org/10.1214/21-AOAS1556SUPPB

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

UNAIDS/WHO (2010). Guidelines on estimating the size of populations most at risk to HIV Technical report, UNAIDS.

WALKER, N., STOVER, J., STANECKI, K., ZANIEWSKI, A. E., GRASSLY, N. C., GARCIA-CALLEJA, J. M. and GHYS, P. D. (2004). The workbook approach to making estimates and projecting future scenarios of HIV/AIDS in countries with low level and concentrated epidemics. *Sex. Transm. Infect.* **80** i10–i13.

ZHENG, T., SALGANIK, M. J. and GELMAN, A. (2006). How many people do you know in prison?: Using overdispersion in count data to estimate social structure in networks. *J. Amer. Statist. Assoc.* **101** 409–423. MR2256163 https://doi.org/10.1198/016214505000001168

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1563–1585
https://doi.org/10.1214/21-AOAS1557
This research was funded, in whole or in part, by [UK Research & Innovation: Economic and Social Research Council, ES/N00812X/1 and ES/S012486/1].

ESTIMATING MODE EFFECTS FROM A SEQUENTIAL MIXED-MODE EXPERIMENT USING STRUCTURAL MOMENT MODELS

BY PAUL S. CLARKE[1,a] AND YANCHUN BAO[2,b]

[1]*Institute for Social & Economic Research, University of Essex,* [a]*pclarke@essex.ac.uk*
[2]*Department of Mathematical Sciences, University of Essex,* [b]*ybaoa@essex.ac.uk*

Until recently, the survey mode of the household panel study *Understanding Society* was mainly face-to-face interview, but it has now adopted a mixed-mode design where individuals can self-complete the questionnaire via the web. As mode is known to affect survey data, a randomized mixed-mode experiment was implemented during the first year of the two-year Wave 8 fieldwork period to assess the impact of this change. The experiment involved a sequential design that permits the identification of mode effects in the presence of nonignorable nonrandom mode selection. While previous studies have used instrumental variables regression to estimate the effects of mode on the means of the survey variables, we describe a more general methodology based on novel structural moment models that characterizes the overall effect of mode on a survey by its effects on the moments of the survey variables' joint distribution. We adapt our estimation procedure to account for nonresponse and complex sampling designs and to include suitable auxiliary data to improve inference and relax key assumptions. Finally, we demonstrate how to estimate the effects of mode on the parameter estimates of generalized linear models and other exponential family models when both outcomes and predictors are subject to mode effects. This methodology is used to investigate the impact of the move to web mode on Wave 8 of *Understanding Society*.

REFERENCES

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. R. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Amer. Statist. Assoc.* **91** 444–472.

BUELENS, B. and VAN DEN BRAKEL, J. A. (2017). Comparing two inferential approaches to handling measurement error in mixed-mode surveys. *J. Off. Stat.* **33** 513–531.

CARPENTER, H. (2018). UK Household Longitudinal Study: Wave 8 Technical Report. Kantar Public. Available at https://doc.ukdataservice.ac.uk/doc/6614/mrdoc/pdf/6614_wave8_technical_report.pdf.

CLARKE, P. S. and BAO, Y. (2022a). Supplement to "Estimating mode effects from a sequential mixed-mode experiment using structural moment models. (Part A)." https://doi.org/10.1214/21-AOAS1557SUPPA

CLARKE, P. S. and BAO, Y. (2022b). Supplement to "Estimating mode effects from a sequential mixed-mode experiment using structural moment models (Part B)." https://doi.org/10.1214/21-AOAS1557SUPPB

CLARKE, P. S., PALMER, T. M. and WINDMEIJER, F. (2015). Estimating structural mean models with multiple instrumental variables using the generalised method of moments. *Statist. Sci.* **30** 96–117. MR3317756 https://doi.org/10.1214/14-STS503

CLARKE, P. S. and WINDMEIJER, F. (2010). Identification of causal effects on binary outcomes using structural mean models. *Biostatistics* **11** 756–770.

D'ARDENNE, J., COLLINS, D., GRAY, M., JESSOP, C. and PILLEY, S. (2017). Assessing the risk of mode effects: Review of proposed survey questions for waves 7-10 of Understanding Society. Understanding Society Working Paper Series 2017-04.

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

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

Key words and phrases. Causal inference, generalized method of moments, encouragement design, instrumental variable, potential outcomes.

HERNÁN, M. A. and ROBINS, J. M. (2020). *Causal Inference*: *What If*. Champman & Hall/CRC, Boca Raton, FL.

IMBENS, G. W. and RUBIN, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models **64** 555–574. MR1485828 https://doi.org/10.2307/2971731

ISER (2018). Understanding Society: Waves 1–8, 2009–2017 and Harmonised BHPS: Waves 1–18, 1991–2009. [data collection], 11th ed. University of Essex, Institute for Social and Economic Research. UK Data Service. SN: 6614. https://doi.org/10.5255/UKDA-SN-6614-13

JÄCKLE, A., GAIA, A. and BENZEVAL, M. (2017). Mixing modes and measurement models in longitudinal studies. CLOSER Resource Report. CLOSER , London, UK.

JÄCKLE, A., ROBERTS, C. and LYNN, P. (2010). Assessing the effect of data collection mode on measurement. *Int*. *Stat*. *Rev.* **78** 3–20.

KOLNENIKOV, S. and KENNEDY, C. (2014). Evaluating three approaches to statistically adjust for mode effects. *Journal of Survey Statistics & Methodology* **2** 126–58.

LUGTIG, P., LENSVELT-MULDERS, G. J. L. M., FRERICHS, R. and GREVEN, A. (2011). Estimating nonresponse bias and mode effects in a mixed-mode survey. *Int*. *J*. *Mark*. *Res*. **53** 669–686.

PARK, S., KIM, J. K. and PARK, S. (2016). An imputation approach for handling mixed-mode surveys. *Ann*. *Appl*. *Stat*. **10** 1063–1085. MR3528372 https://doi.org/10.1214/16-AOAS930

ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm*. *Statist*. *Theory Methods* **23** 2379–2412. MR1293185 https://doi.org/10.1080/03610929408831393

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

TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. Springer, New York. MR2233926

VANNIEUWENHUYZE, J. T. A. (2015). Mode effects on variances, covariances, standard deviations, and correlations. *Journal of Survey Statistics & Methodology* **3** 1–21.

VANNIEUWENHUYZE, J. T. A. and LOOSVELDT, G. (2013). Evaluating relative mode effects in mixed-mode surveys: Three methods to disentangle selection and measurement effects. *Sociol*. *Methods Res*. **42** 82–104. MR3190725 https://doi.org/10.1177/0049124112464868

VANNIEUWENHUYZE, J., LOOSVELDT, G. and MOLENBERGHS, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opin*. *Q*. **74** 1027–1045.

VANNIEUWENHUYZE, J. T. A., LOOSVELDT, G. and MOLENBERGHS, G. (2014). Evaluating mode effects in mixed-mode survey data using covariate adjustment models. *J*. *Off*. *Stat*. **30** 1–21.

VANSTEELANDT, S. and JOFFE, M. (2014). Structural nested models and G-estimation: The partially realized promise. *Statist*. *Sci*. **29** 707–731. MR3300367 https://doi.org/10.1214/14-STS493

WOOLDRIDGE, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press, Cambridge, MA. MR2768559

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1586–1607
https://doi.org/10.1214/21-AOAS1558

MEASURING PERFORMANCE FOR END-OF-LIFE CARE

BY SEBASTIEN HANEUSE[1,a], DEBORAH SCHRAG[2,d], FRANCESCA DOMINICI[1,b], SHARON-LISE NORMAND[3,e] AND KYU HA LEE[1,c]

[1]*Department of Biostatistics, Harvard T.H. Chan School of Public Health,* [a]*shaneuse@hsph.harvard.edu,* [b]*fdominic@hsph.harvard.edu,* [c]*klee@hsph.harvard.edu*

[2]*Division of Population Sciences, Dana-Farber Cancer Institute,* [d]*Deb_Schrag@dfci.harvard.edu*

[3]*Department of Health Care Policy, Harvard Medical School,* [e]*sharon@hcp.med.harvard.edu*

Although not without controversy, readmission is entrenched as a hospital quality metric with statistical analyses generally based on fitting a logistic-Normal generalized linear mixed model. Such analyses, however, ignore death as a competing risk, although doing so for clinical conditions with high mortality can have profound effects; a hospital's seemingly good performance for readmission may be an artifact of it having poor performance for mortality. In this paper we propose novel multivariate hospital-level performance measures for readmission and mortality that derive from framing the analysis as one of cluster-correlated semi-competing risks data. We also consider a number of profiling-related goals, including the identification of extreme performers and a bivariate classification of whether the hospital has higher-/lower-than-expected readmission and mortality rates via a Bayesian decision-theoretic approach that characterizes hospitals on the basis of minimizing the posterior expected loss for an appropriate loss function. In some settings, particularly if the number of hospitals is large, the computational burden may be prohibitive. To resolve this, we propose a series of analysis strategies that will be useful in practice. Throughout, the methods are illustrated with data from CMS on $N = 17{,}685$ patients diagnosed with pancreatic cancer between 2000–2012 at one of $J = 264$ hospitals in California.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A., eds. (1966). *Handbook of Mathematical Functions*, *with Formulas*, *Graphs*, *and Mathematical Tables* **55**. Dover, New York. MR0208797

ALVARES, D., HANEUSE, S., LEE, C. and LEE, K. H. (2019). SemiCompRisks: An R package for the analysis of independent and cluster-correlated semi-competing risks data. *R J.* **11** 376.

ANTONELLI, J., TRIPPA, L. and HANEUSE, S. (2016). Mitigating bias in generalized linear mixed models: The case for Bayesian nonparametrics. *Statist. Sci.* **31** 80–95. MR3458594 https://doi.org/10.1214/15-STS533

BATES, J., LEWIS, S. and PICKARD, A. (2019). *Education Policy*, *Practice and the Professional*. Bloomsbury Publishing.

BUSH, C. A. and MACEACHERN, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83** 275–285.

CHEN, Y., ŞENTÜRK, D., ESTES, J. P., CAMPOS, L. F., RHEE, C. M., DALRYMPLE, L. S., KALANTAR-ZADEH, K. and NGUYEN, D. V. (2021). Performance characteristics of profiling methods and the impact of inadequate case-mix adjustment. *Comm. Statist. Simulation Comput.* **50** 1854–1871. MR4280725 https://doi.org/10.1080/03610918.2019.1595649

CMS (2021). Hospital-Wide All-Cause, Unplanned Readmission Measure (HWR). Available at https://cmit.cms.gov/CMIT_public/ReportMeasure?measureRevisionId=15918. Accessed: 13th July, 2021.

DANIELS, M. J. and NORMAND, S.-L. T. (2005). Longitudinal profiling of health care units based on continuous and discrete patient outcomes. *Biostatistics* **7** 1–15.

DEYO, R. A., CHERKIN, D. C. and CIOL, M. A. (1992). Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J. Clin. Epidemiol.* **45** 613–619.

Key words and phrases. Bayesian decision theory, hierarchical modeling, provider profiling, semicompeting risks, quality of care.

ESTES, J. P., NGUYEN, D. V., CHEN, Y., DALRYMPLE, L. S., RHEE, C. M., KALANTAR-ZADEH, K. and ŞENTÜRK, D. (2018). Time-dynamic profiling with application to hospital readmission among patients on dialysis. *Biometrics* **74** 1383–1394. MR3908156 https://doi.org/10.1111/biom.12908

ESTES, J. P., CHEN, Y., SENTÜRK, D., RHEE, C. M., KÜRÜM, E., YOU, A. S., STREJA, E., KALANTAR-ZADEH, K. and NGUYEN, D. V. (2020). Profiling dialysis facilities for adverse recurrent events. *Stat. Med.* **39** 1374–1389.

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

FINE, J. P. and GRAY, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *J. Amer. Statist. Assoc.* **94** 496–509. MR1702320 https://doi.org/10.2307/2670170

FRIEBEL, R., HAUCK, K., AYLIN, P. and STEVENTON, A. (2018). National trends in emergency readmission rates: A longitudinal analysis of administrative data for England between 2006 and 2016. *BMJ Open* **8** e020325. https://doi.org/10.1136/bmjopen-2017-020325

FRYDMAN, H. and SZAREK, M. (2010). Estimation of overall survival in an 'illness-death' model with application to the vertical transmission of HIV-1. *Stat. Med.* **29** 2045–2054. MR2758446 https://doi.org/10.1002/sim.3949

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

GEORGE, E. I., ROČKOVÁ, V., ROSENBAUM, P. R., SATOPÄÄ, V. A. and SILBER, J. H. (2017). Mortality rate estimation and standardization for public reporting: Medicare's Hospital Compare. *J. Amer. Statist. Assoc.* **112** 933–947. MR3735351 https://doi.org/10.1080/01621459.2016.1276021

GOLDSTEIN, H. and SPIEGELHALTER, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *J. R. Stat. Soc., A* **159** 385–409.

HANEUSE, S., RUDSER, K. D. and GILLEN, D. L. (2008). The separation of timescales in Bayesian survival modeling of the time-varying effect of a time-dependent exposure. *Biostatistics* **9** 400–410.

HANEUSE, S., DOMINICI, F., NORMAND, S.-L. and SCHRAG, D. (2018). Assessment of between-hospital variation in readmission and mortality after cancer surgical procedures. *JAMA Network Open* **1** e183038–e183038.

HANEUSE, S., SCHRAG, D., DOMINICI, F., NORMAND, S.-L. and LEE, K. H. (2022). Supplement to "Measuring performance for end-of-life care." https://doi.org/10.1214/21-AOAS1558SUPPA, https://doi.org/10.1214/21-AOAS1558SUPPB

HATFIELD, L. A., BAUGH, C. M., AZZONE, V. and NORMAND, S.-L. T. (2017). Regulator loss functions and hierarchical modeling for safety decision making. *Med. Decis. Mak.* **37** 512–522.

HE, K., KALBFLEISCH, J. D., LI, Y. and LI, Y. (2013). Evaluating hospital readmission rates in dialysis facilities; adjusting for hospital effects. *Lifetime Data Anal.* **19** 490–512. MR3119994 https://doi.org/10.1007/s10985-013-9264-6

JONES, H. E. and SPIEGELHALTER, D. J. (2011). The identification of "unusual" health-care providers from a hierarchical model. *Amer. Statist.* **65** 154–163. MR2848190 https://doi.org/10.1198/tast.2011.10190

KALBFLEISCH, J. and WOLFE, R. (2013). On monitoring outcomes of medical providers. *Stat. Biosci.* **5** 286–302.

KRISTENSEN, S. R., BECH, M. and QUENTIN, W. (2015). A roadmap for comparing readmission policies with application to Denmark, England, Germany and the United States. *Health Policy* **119** 264–273.

LAIRD, N. and LOUIS, T. (1989). Empirical Bayes ranking methods. *Journal of Educational Statistics* **14** 29–46.

LANDRUM, M. B., NORMAND, S.-L. T. and ROSENHECK, R. A. (2003). Selection of related multivariate means: Monitoring psychiatric care in the Department of Veterans Affairs. *J. Amer. Statist. Assoc.* **98** 7–16. MR1977196 https://doi.org/10.1198/016214503388619049

LECKIE, G. and GOLDSTEIN, H. (2009). The limitations of using school league tables to inform school choice. *J. Roy. Statist. Soc. Ser. A* **172** 835–851. MR2751830 https://doi.org/10.1111/j.1467-985X.2009.00597.x

LEE, K. H., HANEUSE, S., SCHRAG, D. and DOMINICI, F. (2015). Bayesian semiparametric analysis of semicompeting risks data: Investigating hospital readmission after a pancreatic cancer diagnosis. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 253–273. MR3302299 https://doi.org/10.1111/rssc.12078

LEE, K. H., DOMINICI, F., SCHRAG, D. and HANEUSE, S. (2016). Hierarchical models for semicompeting risks data with application to quality of end-of-life care for pancreatic cancer. *J. Amer. Statist. Assoc.* **111** 1075–1095. MR3561930 https://doi.org/10.1080/01621459.2016.1164052

LIN, R., LOUIS, T. A., PADDOCK, S. M. and RIDGEWAY, G. (2006). Loss function based ranking in two-stage, hierarchical models. *Bayesian Anal.* **1** 915–946. MR2282211 https://doi.org/10.1214/06-BA130

LIN, R., LOUIS, T. A., PADDOCK, S. M. and RIDGEWAY, G. (2009). Ranking USRDS provider specific SMRs from 1998-2001. *Health Serv. Outcomes Res. Methodol.* **9** 22–38. https://doi.org/10.1007/s10742-008-0040-0

MAZROUI, Y., MATHOULIN-PELISSIER, S., SOUBEYRAN, P. and RONDEAU, V. (2012). General joint frailty model for recurrent event data with a dependent terminal event: Application to follicular lymphoma data. *Stat. Med.* **31** 1162–1176. MR2925687 https://doi.org/10.1002/sim.4479

McKeague, I. W. and Tighiouart, M. (2000). Bayesian estimators for conditional hazard functions. *Biometrics* **56** 1007–1015. MR1815579 https://doi.org/10.1111/j.0006-341X.2000.01007.x

Millar, R. B. (2009). Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes' factors. *Biometrics* **65** 962–969. MR2649870 https://doi.org/10.1111/j.1541-0420.2008.01162.x

NHS NSS (2019). Hospital Scorecard. Available at https://www.isdscotland.org/Health-Topics/Quality-Indicators/Hospital-Scorecard/. Accessed: 14th April, 2014.

Normand, S.-L. T., Glickman, M. E. and Gatsonis, C. A. (1997). Statistical methods for profiling providers of medical care: Issues and applications. *J. Amer. Statist. Assoc.* **92** 803–814.

Normand, S.-L. T., Ash, A. S., Fienberg, S. E., Stukel, T. A., Utts, J. and Louis, T. A. (2016). League tables for hospital comparisons. *Annu. Rev. Stat. Appl.* **3** 21–50.

Ohlssen, D. I., Sharples, L. D. and Spiegelhalter, D. J. (2007). Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. *Stat. Med.* **26** 2088–2112. MR2364293 https://doi.org/10.1002/sim.2666

Paddock, S. M. (2014). Statistical Benchmarks for Health Care Provider Performance Assessment: A Comparison of Standard Approaches to a Hierarchical Bayesian Histogram-Based Method. *Health Serv. Res.* **49** 1056–1073.

Paddock, S. M. and Louis, T. A. (2011). Percentile-based empirical distribution function estimates for performance evaluation of healthcare providers. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **60** 575–589. MR2829191 https://doi.org/10.1111/j.1467-9876.2010.00760.x

Paddock, S. M., Ridgeway, G., Lin, R. and Louis, T. A. (2006). Flexible distributions of triple-goal estimates in two-stage hierarchical models. *Comput. Statist. Data Anal.* **50** 3243–3262. MR2239666 https://doi.org/10.1016/j.csda.2005.05.008

Ridgeway, G., Nørgaard, M., Rasmussen, T. B., Finkle, W. D., Pedersen, L., Bøtker, H. E. and Sørensen, H. T. (2019). Benchmarking Danish hospitals on mortality and readmission rates after cardiovascular admission. *Clin. Epidemiol.* **11** 67.

Robinson, J. W., Zeger, S. L. and Forrest, C. B. (2006). A hierarchical multivariate two-part model for profiling providers' effects on health care charges. *J. Amer. Statist. Assoc.* **101** 911–923. MR2324092 https://doi.org/10.1198/016214506000000104

Samsky, M. D., Ambrosy, A. P., Youngson, E., Liang, L., Kaul, P., Hernandez, A. F., Peterson, E. D. and McAlister, F. A. (2019). Trends in readmissions and length of stay for patients hospitalized with heart failure in Canada and the United States. *JAMA Cardiol* **4** 444–453. https://doi.org/10.1001/jamacardio.2019.0766

Şentürk, D., Chen, Y., Estes, J. P., Campos, L. F., Rhee, C. M., Kalantar-Zadeh, K. and Nguyen, D. V. (2020). Impact of case-mix measurement error on estimation and inference in profiling of health care providers. *Comm. Statist. Simulation Comput.* **49** 2206–2224. MR4143441 https://doi.org/10.1080/03610918.2018.1515360

Shen, W. and Louis, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 455–471. MR1616061 https://doi.org/10.1111/1467-9868.00135

Silber, J. H., Rosenbaum, P. R., Brachet, T. J., Ross, R. N., Bressler, L. J., Even-Shoshan, O., Lorch, S. A. and Volpp, K. G. (2010). The hospital compare mortality model and the volume–outcome relationship. *Health Serv. Res.* **45** 1148–1167.

Spiegelhalter, D. J. (2005). Funnel plots for comparing institutional performance. *Stat. Med.* **24** 1185–1202. MR2134573 https://doi.org/10.1002/sim.1970

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. MR1979380 https://doi.org/10.1111/1467-9868.00353

Stoer, J. and Bulirsch, R. (2002). *Introduction to Numerical Analysis*, 3rd ed. *Texts in Applied Mathematics* **12**. Springer, New York. MR1923481 https://doi.org/10.1007/978-0-387-21738-3

Varewyck, M., Goetghebeur, E., Eriksson, M. and Vansteelandt, S. (2014). On shrinkage and model extrapolation in the evaluation of clinical center performance. *Biostatistics* **15** 651–664.

Walker, S. G. and Mallick, B. K. (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *J. Roy. Statist. Soc. Ser. B* **59** 845–860. MR1483219 https://doi.org/10.1111/1467-9868.00101

Westert, G. P., Lagoe, R. J., Keskimäki, I., Leyland, A. and Murphy, M. (2002). An international study of hospital readmissions and related utilization in Europe and the USA. *Health Policy* **61** 269–278.

Xu, J., Kalbfleisch, J. D. and Tai, B. (2010). Statistical analysis of illness-death processes and semicompeting risks data. *Biometrics* **66** 716–725. MR2758207 https://doi.org/10.1111/j.1541-0420.2009.01340.x

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1608–1632
https://doi.org/10.1214/21-AOAS1559

SEMIPARAMETRIC MULTINOMIAL MIXED-EFFECTS MODELS: A UNIVERSITY STUDENTS PROFILING TOOL

BY CHIARA MASCI[a], FRANCESCA IEVA[b] AND ANNA MARIA PAGANONI[c]

MOX—Department of Mathematics, Politecnico di Milano, [a]*chiara.masci@polimi.it,* [b]*francesca.ieva@polimi.it,* [c]*anna.paganoni@polimi.it*

Many applicative studies deal with multinomial responses and hierarchical data. Performing clustering at the highest level of grouping, in multilevel multinomial regression, is also often of interest. In this study we analyse Politecnico di Milano data with the aim of profiling students, modelling their probabilities of belonging to different categories and considering their nested structure within engineering degree programmes. In particular, we are interested in clustering degree programmes standing on their effects on different types of student career. To this end, we propose an EM algorithm for implementing semiparametric mixed-effects models dealing with a multinomial response. The novel semiparametric approach assumes the random effects to follow a multivariate discrete distribution with an a priori unknown number of support points, that is, allowed to differ across response categories. The advantage of this modelling is twofold: the discrete distribution on random effects allows, first, to express the marginal density as a weighted sum, avoiding numerical problems in the integration step, typical of the parametric approach, and, second, to identify a latent structure at the highest level of the hierarchy where groups are clustered into subpopulations.

REFERENCES

AGRESTI, A. (2018). *An Introduction to Categorical Data Analysis*. Wiley, New York.

AINA, C. (2013). Parental background and university dropout in Italy. *High. Educ.* **65** 437–456.

AITKIN, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55** 117–128. MR1705676 https://doi.org/10.1111/j.0006-341X.1999.00117.x

ALJOHANI, O. (2016). A comprehensive review of the major studies and theoretical models of student retention in higher education. *High. Educ. Stud.* **6** 1–18.

ANDERSON, D. A. and AITKIN, M. (1985). Variance component models with binary response: Interviewer variability. *J. Roy. Statist. Soc. Ser. B* **47** 203–210. MR0816084

ANDERSON, C. J., KIM, J.-S. and KELLER, B. (2013). Multilevel modeling of categorical response variables. In *Handbook of International Large-Scale Assessment*: *Background*, *Technical Issues*, *and Methods of Data Analysis* 481–519.

ANVUR (2018). Rapporto biennale sullo stato del sistema universitario e della ricerca. Available at https://www.anvur.it/rapporto-biennale/rapporto-biennale-2018.

BARBU, M., VILANOVA, R., VICARIO, J., PEREIRA, M. J., ALVES, P., PODPORA, M., KAWALA-JANIK, A., PRADA, M., DOMINGUEZ, M. et al. (2019). Data mining tool for academic data exploitation: Publication report on engineering students profiles. ERASMUS+ KA2/KA203.

BELLOC, F., MARUOTTI, A. and PETRELLA, L. (2011). How individual characteristics affect university students drop-out: A semiparametric mixed-effects model for an Italian case study. *J. Appl. Stat.* **38** 2225–2239. MR2843254 https://doi.org/10.1080/02664763.2010.545373

BOCK, R. D. and AITKIN, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46** 443–459. MR0668311 https://doi.org/10.1007/BF02293801

BOOTH, J. G. and HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo em algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 265–285.

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

Key words and phrases. Semiparametric statistics, multinomial mixed-effects regression, unsupervised clustering, higher education.

BRESLOW, N. E. and LIN, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82** 81–91. MR1332840 https://doi.org/10.1093/biomet/82.1.81
CANNISTRÀ, M., MASCI, C., IEVA, F., AGASISTI, T. and PAGANONI, A. M. (2021). Early-predicting dropout of university students: an application of innovative machine learning and multilevel statistical techniques *Studies in Higher Education* in press.
COULL, B. A. and AGRESTI, A. (2000). Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics* **56** 73–80. https://doi.org/10.1111/j.0006-341x.2000.00073.x
DE FREITAS, S., GIBSON, D., DU PLESSIS, C., HALLORAN, P., WILLIAMS, E., AMBROSE, M., DUNWELL, I. and ARNAB, S. (2015). Foundations of dynamic learning analytics: Using university student data to increase retention. *Br. J. Educ. Technol.* **46** 1175–1188.
DE LEEUW, J., MEIJER, E. and GOLDSTEIN, H. (2008). *Handbook of Multilevel Analysis*. Springer, Berlin.
DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537
DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR2049007
DOS SANTOS, D. M. and BERRIDGE, D. M. (2000). A continuation ratio random effects model for repeated ordinal responses. *Stat. Med.* **19** 3377–3388.
FONTANA, L., MASCI, C., IEVA, F. and PAGANONI, A. (2021). Performing learning analytics via generalized mixed-effects trees *Data* **6** 7–74.
GOLDSTEIN, H. (2011). *Multilevel Statistical Models* **922**. Wiley, New York.
GOLDSTEIN, H., BROWNE, W. and RASBASH, J. (2002). Partitioning variation in multilevel models. *Underst. Stat.* **1** 223–231.
GOLDSTEIN, H. and RASBASH, J. (1996). Improved approximations for multilevel models with binary responses. *J. Roy. Statist. Soc. Ser. A* **159** 505–513. MR1413664 https://doi.org/10.2307/2983328
HADFIELD, J. D. et al. (2010). Mcmc methods for multi-response generalized linear mixed models: The mcmcglmm R package. *J. Stat. Softw.* **33** 1–22.
HARTZEL, J. S. (2000). Random effects models for nominal and ordinal data.
HARTZEL, J., AGRESTI, A. and CAFFO, B. (2001). Multinomial logit random effects models. *Stat. Model.* **1** 81–102.
HEINEN, T. (1996). *Latent Class and Discrete Latent Trait Models*: *Similarities and Differences*. Sage, Thousand Oaks.
LINDSAY, B. G. (1983a). The geometry of mixture likelihoods: A general theory. *Ann. Statist.* **11** 86–94. MR0684866 https://doi.org/10.1214/aos/1176346059
LINDSAY, B. G. (1983b). The geometry of mixture likelihoods. II. The exponential family. *Ann. Statist.* **11** 783–792. MR0707929 https://doi.org/10.1214/aos/1176346245
MASCI, C., IEVA, F. and PAGANONI, A. M. (2022). Supplement to "Semiparametric multinomial mixed-effects models: A university students profiling tool." https://doi.org/10.1214/21-AOAS1559SUPP
MASCI, C., PAGANONI, A. M. and IEVA, F. (2019). Semiparametric mixed effects models for unsupervised classification of Italian schools. *J. Roy. Statist. Soc. Ser. A* **182** 1313–1342. MR4027363 https://doi.org/10.1111/rssa.12449
MASCI, C., IEVA, F., AGASISTI, T. and PAGANONI, A. M. (2021). Evaluating class and school effects on the joint student achievements in different subjects: A bivariate semiparametric model with random coefficients. *Comput. Statist.* 1–41. https://doi.org/10.1007/s00180-021-01107-1
MCCULLOCH, C. E. (1994). Maximum likelihood variance components estimation for binary data. *J. Amer. Statist. Assoc.* **89** 330–335.
MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92** 162–170. MR1436105 https://doi.org/10.2307/2291460
MCCULLOCH, C. E. and SEARLE, S. R. (2001). *Generalized, Linear, and Mixed Models*. *Wiley Series in Probability and Statistics*: *Texts, References, and Pocketbooks Section*. Wiley-Interscience, New York. MR1884506
MCCULLOCH, C., LIN, H., SLATE, E. and TURNBULL, B. (2002). Discovering subpopulation structure with latent class mixed models. *Stat. Med.* **21** 417–429.
MUTHÉN, B. (2004). Latent variable analysis. *Sage Handb. Quant. Methodol. Soc. Sci.* **345** 106–109.
NAGIN, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychol. Methods* **4** 139.
NAGIN, D. S., JONES, B. L., LIMA PASSOS, V. and TREMBLAY, R. E. (2018). Group-based multi-trajectory modeling. *Stat. Methods Med. Res.* **27** 2015–2023. MR3807918 https://doi.org/10.1177/0962280216673085
PELLAGATTI, M., MASCI, C., IEVA, F. and PAGANONI, A. M. (2021). Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Stat. Anal. Data Min.* **14** 241–257. MR4303069 https://doi.org/10.1002/sam.11505

PINHEIRO, J. and BATES, D. (2006). *Mixed-Effects Models in S and S-PLUS*. Springer, Berlin.
RAUDENBUSH, S. W. (2004). *HLM* 6: *Hierarchical Linear and Nonlinear Modeling*. Scientific Software International.
RAUDENBUSH, S. W., YANG, M.-L. and YOSEF, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *J. Comput. Graph. Statist.* **9** 141–157. MR1826278 https://doi.org/10.2307/1390617
RIGHTS, J. D. and STERBA, S. K. (2016). The relationship between multilevel models and non-parametric multilevel mixture models: Discrete approximation of intraclass correlation, random coefficient distributions, and residual heteroscedasticity. *Br. J. Math. Stat. Psychol.* **69** 316–343.
RODRÍGUEZ, G. and GOLDMAN, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *J. Roy. Statist. Soc. Ser. A* **158** 73–89.
SHAW, D. S., LACOURSE, E. and NAGIN, D. S. (2005). Developmental trajectories of conduct problems and hyperactivity from ages 2 to 10. *J. Child Psychol. Psychiatry* **46** 931–942.
SHAW, D. S., GILLIOM, M., INGOLDSBY, E. M. and NAGIN, D. S. (2003). Trajectories leading to school-age conduct problems. *Dev. Psychol.* **39** 189–200. https://doi.org/10.1037//0012-1649.39.2.189
SKRONDAL, A. and RABE-HESKETH, S. (2004). *Generalized Latent Variable Modeling*: *Multilevel*, *Longitudinal*, *and Structural Equation Models*. *Interdisciplinary Statistics*. CRC Press/CRC, Boca Raton, FL. MR2059021 https://doi.org/10.1201/9780203489437
SPIEGELHALTER, D., THOMAS, A., BEST, N. and LUNN, D. (2003). Winbugs user manual.
STEELE, F., STEELE, F., KALLIS, C., GOLDSTEIN, H. and JOSHI, H. (2005). A multiprocess model for correlated event histories with multiple states, competing risks, and structural effects of one hazard on another. Centre for Multilevel Modelling. http://www.cmm.bristol.ac.uk/research/Multiprocess/mmcehmscrseoha.pdf.
STROUD, A. H. and SECREST, D. (1966). *Gaussian Quadrature Formulas*. Prentice-Hall, Inc., Englewood Cliffs, NJ. MR0202312
R CORE TEAM (2019). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
WOLFINGER, R. and O'CONNELL, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *J. Stat. Comput. Simul.* **48** 233–243.
ZHAO, Y., STAUDENMAYER, J., COULL, B. A. and WAND, M. P. (2006). General design Bayesian generalized linear mixed models. *Statist. Sci.* **21** 35–51. MR2275966 https://doi.org/10.1214/088342306000000015

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1633–1652
https://doi.org/10.1214/21-AOAS1560

CRITICAL WINDOW VARIABLE SELECTION FOR MIXTURES: ESTIMATING THE IMPACT OF MULTIPLE AIR POLLUTANTS ON STILLBIRTH

BY JOSHUA L. WARREN[1,a], HOWARD H. CHANG[2,b], LAUREN K. WARREN[3,c], MATTHEW J. STRICKLAND[4,d], LYNDSEY A. DARROW[4,e] AND JAMES A. MULHOLLAND[5,f]

[1]*Department of Biostatistics, Yale University,* [a]*joshua.warren@yale.edu*
[2]*Department of Biostatistics and Bioninformatics, Emory University,* [b]*howard.chang@emory.edu*
[3]*RTI International,* [c]*lklein@rti.org*
[4]*School of Public Health, University of Nevada,* [d]*mstrickland@unr.edu,* [e]*ldarrow@unr.edu*
[5]*School of Civil and Environmental Engineering, Georgia Institute of Technology,* [f]*james.mulholland@ce.gatech.edu*

Understanding the role of time-varying pollution mixtures on human health is critical as people are simultaneously exposed to multiple pollutants during their lives. For vulnerable subpopulations who have well-defined exposure periods (e.g., pregnant women), questions regarding critical windows of exposure to these mixtures are important for mitigating harm. We extend critical window variable selection (CWVS) to the multipollutant setting by introducing CWVS for mixtures (CWVSmix), a hierarchical Bayesian method that combines smoothed variable selection and temporally correlated weight parameters to: (i) identify critical windows of exposure to mixtures of time-varying pollutants, (ii) estimate the time-varying relative importance of each individual pollutant and their first order interactions within the mixture, and (iii) quantify the impact of the mixtures on health. Through simulation we show that CWVSmix offers the best balance of performance in each of these categories in comparison to competing methods. Using these approaches, we investigate the impact of exposure to multiple ambient air pollutants on the risk of stillbirth in New Jersey, 2005–2014. We find consistent elevated risk in gestational weeks 2, 16–17, and 20 for non-Hispanic Black mothers, with pollution mixtures dominated by ammonium (weeks 2, 17, 20), nitrate (weeks 2, 17), nitrogen oxides (weeks 2, 16), $PM_{2.5}$ (week 2), and sulfate (week 20). The method is available in the R package `CWVSmix`.

REFERENCES

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

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

BEKKAR, B., PACHECO, S., BASU, R. and DENICOLA, N. (2020). Association of air pollution and heat exposure with preterm birth, low birth weight, and stillbirth in the US: A systematic review. *JAMA Netw. Open* **3** e208243. https://doi.org/10.1001/jamanetworkopen.2020.8243

BELLO, G. A., ARORA, M., AUSTIN, C., HORTON, M. K., WRIGHT, R. O. and GENNINGS, C. (2017). Extending the distributed lag model framework to handle chemical mixtures. *Environ. Res.* **156** 253–264.

BRUNEKREEF, B. and HOLGATE, S. T. (2002). Air pollution and health. *Lancet* **360** 1233–1242.

CARRICO, C., GENNINGS, C., WHEELER, D. C. and FACTOR-LITVAK, P. (2015). Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *J. Agric. Biol. Environ. Stat.* **20** 100–120. MR3334469 https://doi.org/10.1007/s13253-014-0180-3

Key words and phrases. Bayesian variable selection, correlated weights, multivariate latent variables, pollution mixtures, reproductive health.

CHANG, H. H., WARREN, J. L., DARROW, L. A., REICH, B. J. and WALLER, L. A. (2015). Assessment of critical exposure and outcome windows in time-to-event analysis with application to air pollution and preterm birth study. *Biostatistics* **16** 509–521. MR3365443 https://doi.org/10.1093/biostatistics/kxu060

CHEN, Y.-H., MUKHERJEE, B. and BERROCAL, V. J. (2019). Distributed lag interaction models with two pollutants. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 79–97. MR3902983

COLICINO, E., PEDRETTI, N. F., BUSGANG, S. A. and GENNINGS, C. (2020). Per-and poly-fluoroalkyl substances and bone mineral density: Results from the Bayesian weighted quantile sum regression. *Environ. Epidemiol.* **4**.

DAVALOS, A. D., LUBEN, T. J., HERRING, A. H. and SACKS, J. D. (2017). Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures. *Ann. Epidemiol.* **27** 145–153.

DOMINICI, F., PENG, R. D., BARR, C. D. and BELL, M. L. (2010). Protecting human health from air pollution: Shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology* **21** 187.

FAIZ, A. S., RHOADS, G. G., DEMISSIE, K., KRUSE, L., LIN, Y. and RICH, D. Q. (2012). Ambient air pollution and the risk of stillbirth. *Am. J. Epidemiol.* **176** 308–316.

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

FERRARI, F. and DUNSON, D. B. (2021). Bayesian Factor Analysis for Inference on Interactions. *J. Amer. Statist. Assoc.* **116** 1521–1532. MR4309290 https://doi.org/10.1080/01621459.2020.1745813

GENNINGS, C., CURTIN, P., BELLO, G., WRIGHT, R., ARORA, M. and AUSTIN, C. (2020). Lagged WQS regression for mixtures with many components. *Environ. Res.* **86** 109529.

GEWEKE, J. et al. (1991). *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments* **196**. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, Minneapolis, MN, USA.

GREEN, R., SAROVAR, V., MALIG, B. and BASU, R. (2015). Association of stillbirth with ambient air pollution in a California cohort study. *Am. J. Epidemiol.* **181** 874–882.

HORTON, R. and SAMARASEKERA, U. (2016). Stillbirths: Ending an epidemic of grief. *Lancet* **387** 515–516. https://doi.org/10.1016/S0140-6736(15)01276-3

HOYERT, D. L. and GREGORY, E. C. W. (2016). Cause of fetal death: Data from the fetal death report, 2014. *Natl. Vital Stat. Rep.* **65** 1–25.

KAMPA, M. and CASTANAS, E. (2008). Human health effects of air pollution. *Environ. Pollut.* **151** 362–367.

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, S. H., BOBB, J. F., LEE, K. H. et al. (2018a). Lagged kernel machine regression for identifying time windows of susceptibility to exposures of complex mixtures. *Biostatistics* **19** 325–341. MR3815175 https://doi.org/10.1093/biostatistics/kxx036

LIU, S. H., BOBB, J. F., CLAUS HENN, B., SCHNAAS, L., TELLEZ-ROJO, M. M., GENNINGS, C., ARORA, M., WRIGHT, R. O., COULL, B. A. et al. (2018b). Modeling the health effects of time-varying complex environmental mixtures: Mean field variational Bayes for lagged kernel machine regression. *Environmetrics* **29** e2504. MR3813303 https://doi.org/10.1002/env.2504

POPE, D. P., MISHRA, V., THOMPSON, L., SIDDIQUI, A. R., REHFUESS, E. A., WEBER, M. and BRUCE, N. G. (2010). Risk of low birth weight and stillbirth associated with indoor air pollution from solid fuel use in developing countries. *Epidemiol. Rev.* **32** 70–81. https://doi.org/10.1093/epirev/mxq005

REICH, B. J., GUAN, Y., FOURCHES, D., WARREN, J. L., SARNAT, S. E. and CHANG, H. H. (2020). Integrative statistical methods for exposure mixtures and health. *Ann. Appl. Stat.* **14** 1945–1963. MR4194255 https://doi.org/10.1214/20-AOAS1364

RENZETTI, S., CURTIN, P., JUST, A. C., BELLO, G. and GENNINGS, C. (2020). gWQS: Generalized Weighted Quantile Sum Regression. R package version 3.0.0.

SENTHILKUMAR, N., GILFETHER, M., METCALF, F., RUSSELL, A. G., MULHOLLAND, J. A. and CHANG, H. H. (2019). Application of a fusion method for gas and particle air pollutants between observational data and chemical transport model simulations over the contiguous United States for 2005–2014. *Int. J. Environ. Res. Public Health* **16** 3314.

SIDDIKA, N., BALOGUN, H. A., AMEGAH, A. K. and JAAKKOLA, J. J. (2016). Prenatal ambient air pollution exposure and the risk of stillbirth: Systematic review and meta-analysis of the empirical evidence. *Occup. Environ. Med.* **73** 573–581.

STIEB, D. M., CHEN, L., ESHOUL, M. and JUDEK, S. (2012). Ambient air pollution, birth weight and preterm birth: A systematic review and meta-analysis. *Environ. Res.* **117** 100–111.

STRAND, L. B., BARNETT, A. G. and TONG, S. (2011). Methodological challenges when estimating the effects of season and seasonal exposures on birth outcomes. *BMC Med. Res. Methodol.* **11** 49. https://doi.org/10.1186/1471-2288-11-49

WACKERNAGEL, H. (2013). *Multivariate Geostatistics*: *An Introduction with Applications*. Springer, Berlin.

WARREN, J. L., LUBEN, T. J. and CHANG, H. H. (2020). A spatially varying distributed lag model with application to an air pollution and term low birth weight study. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 681–696. MR4098968

WARREN, J., FUENTES, M., HERRING, A. and LANGLOIS, P. (2012a). Bayesian spatial-temporal model for cardiac congenital anomalies and ambient air pollution risk assessment. *Environmetrics* **23** 673–684. MR3019059 https://doi.org/10.1002/env.2174

WARREN, J., FUENTES, M., HERRING, A. and LANGLOIS, P. (2012b). Spatial-temporal modeling of the association between air pollution exposure and preterm birth: Identifying critical windows of exposure. *Biometrics* **68** 1157–1167. MR3040022 https://doi.org/10.1111/j.1541-0420.2012.01774.x

WARREN, J. L., STINGONE, J. A., HERRING, A. H. et al. (2016). Bayesian multinomial probit modeling of daily windows of susceptibility for maternal $PM_{2.5}$ exposure and congenital heart defects. *Stat. Med.* **35** 2786–2801. MR3513718 https://doi.org/10.1002/sim.6891

WARREN, J. L., SON, J.-Y., PEREIRA, G., LEADERER, B. P. and BELL, M. L. (2017). Investigating the impact of maternal residential mobility on identifying critical windows of susceptibility to ambient air pollution during pregnancy. *Am. J. Epidemiol.* **187** 992–1000.

WARREN, J. L., KONG, W., LUBEN, T. J. and CHANG, H. H. (2020). Critical window variable selection: Estimating the impact of air pollution on very preterm birth. *Biostatistics* **21** 790–806. MR4164058 https://doi.org/10.1093/biostatistics/kxz006

WARREN, J. L., CHANG, H. H., WARREN, L. K., STRICKLAND, M. J., DARROW, L. A. and MULHOLLAND, J. A. (2022). Supplement to "Critical window variable selection for mixtures: Estimating the impact of multiple air pollutants on stillbirth." https://doi.org/10.1214/21-AOAS1560SUPPA, https://doi.org/10.1214/21-AOAS1560SUPPB

WILSON, A., CHIU, Y.-H. M., HSU, H.-H. L., WRIGHT, R. O., WRIGHT, R. J. and COULL, B. A. (2017). Bayesian distributed lag interaction models to identify perinatal windows of vulnerability in children's health. *Biostatistics* **18** 537–552. MR3799593 https://doi.org/10.1093/biostatistics/kxx002

WILSON, A., HSU, H.-H. L., CHIU, Y.-H. M., WRIGHT, R. O., WRIGHT, R. J. and COULL, B. A. (2019). Kernel machine and distributed lag models for assessing windows of susceptibility to mixtures of time-varying environmental exposures in children's health studies. ArXiv Preprint. Available at arXiv:1904.12417.

ZHANG, H., ZHANG, X., WANG, Q., XU, Y., FENG, Y., YU, Z. and HUANG, C. (2021). Ambient air pollution and stillbirth: An updated systematic review and meta-analysis of epidemiological studies. *Environ. Pollut.* 116752.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1653–1675
https://doi.org/10.1214/21-AOAS1561

HIGH-RESOLUTION BAYESIAN MAPPING OF LANDSLIDE HAZARD WITH UNOBSERVED TRIGGER EVENT

BY THOMAS OPITZ[1,a], HAAKON BAKKA[2,b], RAPHAËL HUSER[3,c] AND LUIGI LOMBARDO[4,d]

[1]*Biostatistics and Spatial Processes, INRAE,* [a]*thomas.opitz@inrae.fr*

[2]*Department of Mathematics, University of Oslo,* [b]*haakoncb@math.uio.no*

[3]*Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST),* [c]*raphael.huser@kaust.edu.sa*

[4]*Faculty of Geo-Information Science and Earth Observation, University of Twente,* [d]*l.lombardo@utwente.nl*

Statistical models for landslide hazard enable mapping of risk factors and landslide occurrence intensity by using geomorphological covariates available at high spatial resolution. However, the spatial distribution of the triggering event (e.g., precipitation or earthquakes) is often not directly observed. In this paper we develop Bayesian spatial hierarchical models for point patterns of landslide occurrences using different types of log-Gaussian Cox processes. Starting from a competitive baseline model that captures the unobserved precipitation trigger through a spatial random effect at slope unit resolution, we explore novel complex model structures that take clusters of events arising at small spatial scales into account as well as nonlinear or spatially-varying covariate effects. For a 2009 event of around 5000 precipitation-triggered landslides in Sicily, Italy, we show how to fit our proposed models efficiently, using the integrated nested Laplace approximation (INLA), and rigorously compare the performance of our models both from a statistical and applied perspective. In this context we argue that model comparison should not be based on a single criterion and that different models of various complexity may provide insights into complementary aspects of the same applied problem. In our application our models are found to have mostly the same spatial predictive performance, implying that key to successful prediction is the inclusion of a slope-unit resolved random effect capturing the precipitation trigger. Interestingly, a parsimonious formulation of space-varying slope effects reflects a physical interpretation of the precipitation trigger: in subareas with weak trigger, the slope steepness is shown to be mostly irrelevant.

REFERENCES

ALVIOLI, M., MARCHESINI, I., REICHENBACH, P., ROSSI, M., ARDIZZONE, F., FIORUCCI, F. and GUZZETTI, F. (2016). Automatic delineation of geomorphological slope units with r.slopeunits v1.0 and their optimization for landslide susceptibility modeling. *Geosci. Model Dev.* **9** 3975–3991.

AMATO, G., EISANK, C., CASTRO-CAMILO, D. and LOMBARDO, L. (2019). Accounting for covariate distributions in slope-unit-based landslide susceptibility models. A case study in the Alpine environment. *Eng. Geol.* **260** 105237.

ARNONE, E., FRANCIPANE, A., SCARBACI, A., PUGLISI, C. and NOTO, L. (2016). Effect of raster resolution and polygon-conversion algorithm on landslide susceptibility mapping. *Environ. Model. Softw.* **84** 467–481.

ATKINSON, P. M. and MASSARI, R. (1998). Generalised linear modelling of susceptibility to landsliding in the central Apennines, Italy. *Comput. Geosci.* **24** 373–385.

AYALEW, L. and YAMAGISHI, H. (2005). The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda–Yahiko Mountains, Central Japan. *Geomorphology* **65** 15–31.

BESAG, J. (1975). Statistical analysis of non-lattice data. *J. Roy. Stat. Soc.* (*Ser. D*) 179–195.

BEVEN, K. and KIRKBY, M. J. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrol. Sci. J.* **24** 43–69.

Key words and phrases. Integrated nested Laplace approximation (INLA), landslide, log-Gaussian Cox process, model complexity, space-varying regression, triggering event.

BOUT, B., LOMBARDO, L., VAN WESTEN, C. J. and JETTEN, V. G. (2018). Integration of two-phase solid fluid equations in a catchment model for flashfloods, debris flows and shallow slope failures. *Environ. Model. Softw.* **105** 1–16.

CAMA, M., LOMBARDO, L., CONOSCENTI, C., AGNESI, V. and ROTIGLIANO, E. (2015). Predicting storm-triggered debris flow events: Application to the 2009 Ionian Peloritan disaster (Sicily, Italy). *Nat. Hazards Earth Syst. Sci.* **15** 1785–1806.

CAMA, M., CONOSCENTI, C., LOMBARDO, L. and ROTIGLIANO, E. (2016). Exploring relationships between grid cell size and accuracy for debris-flow susceptibility models: A test in the Giampilieri catchment (Sicily, Italy). *Environmental Earth Sciences* **75** 1–21.

CARRARA, A., CARDINALI, M., GUZZETTI, F. and REICHENBACH, P. (1995). GIS technology in mapping landslide hazard. In *Geographical Information Systems in Assessing Natural Hazards* 135–175. Springer, Berlin.

CASTRO CAMILO, D., LOMBARDO, L., MAI, P. M., DOU, J. and HUSER, R. (2017). Handling high predictor dimensionality in slope-unit-based landslide susceptibility models through LASSO-penalized generalized linear model. *Environ. Model. Softw.* **97** 145–156.

COROMINAS, J., VAN WESTEN, C., FRATTINI, P., CASCINI, L., MALET, J.-P., FOTOPOULOU, S., CATANI, F., VAN DEN EECKHAUT, M., MAVROULI, O. et al. (2014). Recommendations for the quantitative analysis of landslide risk. *Bulletin of Engineering Geology and the Environment* **73** 209–263.

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

GAMERMAN, D., MOREIRA, A. R. B. and RUE, H. (2003). Space-varying regression models: Specifications and simulation. *Comput. Statist. Data Anal.* **42** 513–533. MR2005406 https://doi.org/10.1016/S0167-9473(02)00211-6

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

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

GNEITING, T. and KATZFUSS, M. (2014). Probabilistic forecasting. *Annu. Rev. Stat. Appl.* **1** 125–151.

GOETZ, J., BRENNING, A., PETSCHKO, H. and LEOPOLD, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Comput. Geosci.* **81** 1–11.

GUZZETTI, F. and REICHENBACH, P. (1994). Towards a definition of topographic divisions for Italy. *Geomorphology* **11** 57–74.

HEERDEGEN, R. G. and BERAN, M. A. (1982). Quantifying source areas through land surface curvature and shape. *J. Hydrol.* **57** 359–373.

HUNGR, O., LEROUEIL, S. and PICARELLI, L. (2014). The varnes classification of landslide types, an update. *Landslides* **11** 167–194.

ILLIAN, J. B., SØRBYE, S. H. and RUE, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *Ann. Appl. Stat.* **6** 1499–1530. MR3058673 https://doi.org/10.1214/11-AOAS530

KOH, J., PIMONT, F., DUPUY, J.-L. and OPITZ, T. (2021). Spatiotemporal wildfire modeling through point processes with moderate and extreme marks. ArXiv Preprint. Available at arXiv:2105.08004.

KRAINSKI, E. T., GÓMEZ-RUBIO, V., BAKKA, H., LENZI, A., CASTRO-CAMILO, D., SIMPSON, D., LINDGREN, F. and RUE, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. CRC Press/CRC, New York.

LEININGER, T. J. and GELFAND, A. E. (2017). Bayesian inference and model assessment for spatial point patterns using posterior predictive samples. *Bayesian Anal.* **12** 1–30. MR3597565 https://doi.org/10.1214/15-BA985

LOMBARDO, L., OPITZ, T. and HUSER, R. (2018). Point process-based modeling of multiple debris flow landslides using INLA: An application to the 2009 Messina disaster. *Stoch. Environ. Res. Risk Assess.* **32** 2179–2198.

LOMBARDO, L., OPITZ, T. and HUSER, R. (2019). Numerical recipes for landslide spatial prediction using R-INLA: A step-by-step tutorial. In *Spatial Modeling in GIS and R for Earth and Environmental Sciences* (H. R. Pourghasemi and C. Gokceoglu, eds.) 55–83. Elsevier, Amsterdam.

LOMBARDO, L., CAMA, M., MAERKER, M. and ROTIGLIANO, E. (2014). A test of transferability for landslides susceptibility models under extreme climatic events: Application to the Messina 2009 disaster. *Natural Hazards* **74** 1951–1989.

LOMBARDO, L., FUBELLI, G., AMATO, G. and BONASERA, M. (2016a). Presence-only approach to assess landslide triggering-thickness susceptibility: A test for the Mili catchment (North-Eastern Sicily, Italy). *Natural Hazards* **84** 565–588.

LOMBARDO, L., BACHOFER, F., CAMA, M., MÄRKER, M. and ROTIGLIANO, E. (2016b). Exploiting maximum entropy method and ASTER data for assessing debris flow and debris slide susceptibility for the Giampilieri catchment (North-Eastern Sicily, Italy). *Earth Surf. Process. Landf.* **41** 1776–1789.

LOMBARDO, L., BAKKA, H., TANYAS, H., VAN WESTEN, C., MAI, P. M. and HUSER, R. (2019). Geostatistical modeling to capture seismic-shaking patterns from earthquake-induced landslides. *J. Geophys. Res., Earth Surf.* **124** 1958–1980.

LOMBARDO, L., OPITZ, T., ARDIZZONE, F., GUZZETTI, F. and HUSER, R. (2020). Space-time landslide predictive modelling. *Earth-Sci. Rev.* 103318.

MØLLER, J., SYVERSVEEN, A. R. and WAAGEPETERSEN, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Stat.* **25** 451–482. MR1650019 https://doi.org/10.1111/1467-9469.00115

MOORE, I. D., GRAYSON, R. and LADSON, A. (1991). Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrol. Process.* **5** 3–30.

MORAGA, P. (2019). *Geospatial Health Data*: *Modeling and Visualization with R-INLA and Shiny*. CRC Press/CRC Biostatistics Series, Boca Raton, FL.

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

OPITZ, T. (2017). Latent Gaussian modeling and INLA: A review with focus on space-time applications. *J. French Stat. Soc.* **158** 62–85. MR3720130

OPITZ, T., HUSER, R., BAKKA, H. and RUE, H. (2018). INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. *Extremes* **21** 441–462. MR3855716 https://doi.org/10.1007/s10687-018-0324-x

OPITZ, T., BAKKA, H., HUSER, R. and LOMBARDO, L. (2022). Supplement to "High-resolution Bayesian mapping of landslide hazard with unobserved trigger event." https://doi.org/10.1214/21-AOAS1561SUPP

REICHENBACH, P., ROSSI, M., MALAMUD, B. D., MIHIR, M. and GUZZETTI, F. (2018). A review of statistically-based landslide susceptibility models. *Earth-Sci. Rev.* **180** 60–91.

ROSSI, M., GUZZETTI, F., REICHENBACH, P., MONDINI, A. C. and PERUCCACCI, S. (2010). Optimal landslide susceptibility zonation based on multiple forecasts. *Geomorphology* **114** 129–142.

ROUSE JR., J., HAAS, R., SCHELL, J. and DEERING, D. (1974). Monitoring vegetation systems in the Great Plains with ERTS.

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

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

RUE, H., RIEBLER, A., SØRBYE, S. H., ILLIAN, J. B., SIMPSON, D. P. and LINDGREN, F. K. (2016). Bayesian computing with INLA: A review. *Annu. Rev. Stat. Appl.* **1**.

SIMPSON, D., RUE, H., RIEBLER, A., MARTINS, T. G. and SØRBYE, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.* **32** 1–28. MR3634300 https://doi.org/10.1214/16-STS576

SØRBYE, S. H. and RUE, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spat. Stat.* **8** 39–51. MR3326820 https://doi.org/10.1016/j.spasta.2013.06.004

TENG, M., NATHOO, F. and JOHNSON, T. D. (2017). Bayesian computation for Log-Gaussian Cox processes: A comparative analysis of methods. *J. Stat. Comput. Simul.* **87** 2227–2252. MR3656102 https://doi.org/10.1080/00949655.2017.1326117

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. MR0830567

VAN DEN BOUT, B., LOMBARDO, L., CHIYANG, M., VAN WESTEN, C. and JETTEN, V. (2021). Physically-based catchment-scale prediction of slope failure volume and geometry. *Eng. Geol.* 105942.

VARNES, D. J. (1958). Landslide types and processes. *Landslides and Engineering Practice* **24** 20–47.

VRANCKX, M., NEYENS, T. and FAES, C. (2019). Comparison of different software implementations for spatial disease mapping. *Spat. Spatiotemporal. Epidemiol.* **31** 100302. https://doi.org/10.1016/j.sste.2019.100302

WILSON, J. P. and GALLANT, J. C. (2000). Digital terrain analysis. *Terrain Analysis*: *Principles and Applications* **6** 1–27.

ZEVENBERGEN, L. W. and THORNE, C. R. (1987). Quantitative analysis of land surface topography. *Earth Surf. Process. Landf.* **12** 47–56.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1676–1699
https://doi.org/10.1214/21-AOAS1562

BAYESIAN FUNCTIONAL REGISTRATION OF FMRI ACTIVATION MAPS

BY GUOQING WANG[a], ABHIRUP DATTA[b] AND MARTIN A. LINDQUIST[c]

Department of Biostatistics, Johns Hopkins University, [a]*gqw@jhu.edu,* [b]*abhidatta@jhu.edu,* [c]*mlindquist@jhu.edu*

Functional magnetic resonance imaging (fMRI) has provided invaluable insight into our understanding of human behavior. However, large interindividual differences in both brain anatomy and functional localization *after* anatomical alignment remain a major limitation in conducting group analyses and performing population level inference. This paper addresses this problem by developing and validating a new computational technique for reducing misalignment across individuals in functional brain systems by spatially transforming each subject's functional data to a common reference map. Our proposed Bayesian functional registration approach allows us to assess differences in brain function across subjects and individual differences in activation topology. It combines intensity-based and feature-based information into an integrated framework and allows inference to be performed on the transformation via the posterior samples. We evaluate the method in a simulation study and apply it to data from a study of thermal pain. We find that the proposed approach provides increased sensitivity for group-level inference.

REFERENCES

ALLISON, T., PUCE, A., SPENCER, D. D. and MCCARTHY, G. (1999). Electrophysiological studies of human face perception. I: Potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cereb. Cortex* **9** 415–430.

AMUNTS, K., MALIKOVIC, A., MOHLBERG, H., SCHORMANN, T. and ZILLES, K. (2000). Brodmann's areas 17 and 18 brought into stereotaxic space-where and how variable? *NeuroImage* **11** 66–84. https://doi.org/10.1006/nimg.1999.0516

AWANGE, J. L., BAE, K. H. and CLAESSENS, S. J. (2008). Procrustean solution of the 9-parameter transformation problem. *Earth Planets Space* **60** 529–537.

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, Second Edition ed. CRC Press, Philadelphia, PA.

BESL, P. J. and MCKAY, N. D. (1992). A method for registration of 3-D shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **14** 239–256.

BISSIRI, P. G., HOLMES, C. C. and WALKER, S. G. (2016). A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 1103–1130. MR3557191 https://doi.org/10.1111/rssb.12158

BRADLEY, D. and ROTH, G. (2007). Adaptive thresholding using the integral image. *J. Graphics Tools* **12** 13–21.

BRUNET, F., BARTOLI, A., NAVAB, N. and MALGOUYRES, R. (2010). Pixel-based hyperparameter selection for feature-based image registration. *VMV* 2010—*Vision. Model. Vis.* 33–40.

BUSHNELL, M., DUNCAN, G., HOFBAUER, R., HA, B., CHEN, J.-I. and CARRIER, B. (1999). Pain perception: Is there a role for primary somatosensory cortex? *Proc. Natl. Acad. Sci. USA* **96** 7705–7709.

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

CEDERBERG, J. N. (2001). *A Course in Modern Geometries*, 2nd ed. *Undergraduate Texts in Mathematics*. Springer, New York. MR1798736 https://doi.org/10.1007/978-1-4757-3490-4

CHEN, P. H., GUNTUPALLI, J. S., HAXBY, J. V. and RAMADGE, P. J. (2014). Joint SVD-hyperalignment for multi-subject FMRI data alignment. In *IEEE International Workshop on Machine Learning for Signal Processing*, *MLSP* 1–6.

CHERNOZHUKOV, V. and HONG, H. (2003). An MCMC approach to classical estimation. *J. Econometrics* **115** 293–346. MR1984779 https://doi.org/10.1016/S0304-4076(03)00100-3

Key words and phrases. Functional magnetic resonance imaging, group-level analysis, registration, Bayesian methods, interindividual differences, pain.

CHRISTENSEN, G. and JOHNSON, H. (2001). Consistent image registration. *IEEE Trans. Med. Imag.* **20** 568–582.

CHUMCHOB, N. and CHEN, K. (2009). A robust affine image registration method. *Int. J. Numer. Anal. Model.* **6** 311–334. MR2574911

CONROY, B. R., SINGER, B. D., GUNTUPALLI, J. S., RAMADGE, P. J. and HAXBY, J. V. (2013). Inter-subject alignment of human cortical anatomy using functional connectivity. *NeuroImage* **81** 400–411. https://doi.org/10.1016/j.neuroimage.2013.05.009

CRESSIE, N. and WIKLE, C. K. (2015). *Statistics for Spatio-Temporal Data*. Wiley, New York.

CRUM, W. R., HARTKENS, T. and HILL, D. L. (2004). Non-rigid image registration: Theory and practice. *Br. J. Radiol.* **77**.

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

DUNCAN, K. J., PATTAMADILOK, C., KNIERIM, I. and DEVLIN, J. T. (2009). Consistency and variability in functional localisers. *NeuroImage* **46** 1018–1026.

ESTER, M., KRIEGEL, H.-P., SANDER, J. and XU, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* 226–231.

FIKSEL, J., DATTA, A., AMOUZOU, A. and ZEGER, S. (2021). Generalized Bayes quantification learning under dataset shift. *J. Amer. Statist. Assoc.* 1–19.

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

FISCHER, B. and MODERSITZKI, J. (2003). Curvature based image registration. *J. Math. Imaging Vision* **18** 81–85.

FISCHER, B. and MODERSITZKI, J. (2008). Ill-posed medicine—an introduction to image registration. *Inverse Probl.* **24** 034008. MR2421945 https://doi.org/10.1088/0266-5611/24/3/034008

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences linked references are available on JSTOR for this article: Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GRUEN, A. W. and BALTSAVIAS, E. P. (1987). High-precision image matching for digital terrain model generation. *Photogrammetria* **42** 97–112.

GRÜNWALD, P. D. and MEHTA, N. A. (2020). Fast rates for general unbounded loss functions: From ERM to generalized Bayes. *J. Mach. Learn. Res.* **21** 56. MR4095335

GUNTUPALLI, J. S., HANKE, M., HALCHENKO, Y. O., CONNOLLY, A. C., RAMADGE, P. J. and HAXBY, J. V. (2016). A model of representational spaces in human cortex. *Cereb. Cortex* **26** 2919–2934.

GUSTIN, S. M., PECK, C. C., CHENEY, L. B., MACEY, P. M., MURRAY, G. M. and HENDERSON, L. A. (2012). Pain and plasticity: Is chronic pain always associated with somatosensory cortex activity and reorganization? *J. Neurosci.* **32** 14874–14884.

HASSON, U., NIR, Y., LEVY, I., FUHRMANN, G. and MALACH, R. (2004). Intersubject synchronization of cortical activity during. *Nat. Vis. Sci.* **303** 1634–1640.

HAXBY, J. V., GUNTUPALLI, J. S., CONNOLLY, A. C., HALCHENKO, Y. O., CONROY, B. R., GOBBINI, M. I., HANKE, M. and RAMADGE, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72** 404–416.

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

IORDAN, M. C., JOULIN, A., BECK, D. M. and FEI-FEI, L. (2016). Locally-optimized inter-subject alignment of functional cortical regions. arXiv preprint. Available at arXiv:1606.02349.

JOHNSON, H. and CHRISTENSEN, G. (2002). Consistent landmark and intensity-based image registration. *IEEE Trans. Med. Imag.* **21** 450–461.

LINDQUIST, M. A. (2008). The statistical analysis of fMRI data. *Statist. Sci.* **23** 439–464. MR2530545 https://doi.org/10.1214/09-STS282

LINDQUIST, M. A., LOH, J. M., ATLAS, L. Y. and WAGER, T. D. (2009). Modeling the hemodynamic response function in fmri: Efficiency, bias and mis-modeling. *NeuroImage* **45** S187–S198.

LORBERT, A. and RAMADGE, P. J. (2012). Kernel hyperalignment. *Adv. Neural Inf. Process. Syst.* **3** 1790–1798.

MATHERON, G. (1963). Principles of geostatistics. *Econ. Geol.* **58** 1246–1266.

MCALLESTER, D. A. (1999). Some pac-Bayesian theorems. *Machine Learning* **37** 355–363.

MCCARTHY, G., PUCE, A., BELGER, A. and ALLISON, T. (1999). Electrophysiological studies of human face perception. II: Response properties of face-specific potentials generated in occipitotemporal cortex. *Cereb. Cortex* **9** 431–444.

MIAN, A. S., BENNAMOUN, M. and OWENS, R. (2006). Three-dimensional model-based object recognition and segmentation in cluttered scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **28** 1584–1601.

NENNING, K. H., LIU, H., GHOSH, S. S., SABUNCU, M. R., SCHWARTZ, E. and LANGS, G. (2017). Diffeomorphic functional brain surface alignment: Functional demons. *NeuroImage* **156** 456–465.

OMBAO, H., LINDQUIST, M., THOMPSON, W. and ASTON, J. (2016). *Handbook of Neuroimaging Data Analysis*. CRC Press, Boca Raton.

RADEMACHER, J., CAVINESS, V. JR, STEINMETZ, H. and GALABURDA, A. (1993). Topographical variation of the human primary cortices: Implications for neuroimaging, brain mapping, and neurobiology. *Cereb. Cortex* **3** 313–329.

RIGON, T., HERRING, A. H. and DUNSON, D. B. (2020). A generalized bayes framework for probabilistic clustering. arXiv preprint. Available at arXiv:2006.05451.

SABUNCU, M. R., SINGER, B. D., CONROY, B., BRYAN, R. E., RAMADGE, P. J. and HAXBY, J. V. (2010). Function-based intersubject alignment of human cortical anatomy. *Cereb. Cortex* **20** 130–140.

SHAWE-TAYLOR, J. and WILLIAMSON, R. C. (1997). A pac analysis of a Bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory* 2–9.

SIMARD, P., STEINKRAUS, D. and PLATT, J. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition*, 2003. *Proceedings* 958–963.

SIMPSON, I. J., SCHNABEL, J. A., GROVES, A. R., ANDERSSON, J. L. and WOOLRICH, M. W. (2012). Probabilistic inference of regularisation in non-rigid registration. *NeuroImage* **59** 2438–2451.

THEVENAZ, P. and UNSER, M. (1998). Efficient mutual information optimizer for multiresolution image registration. *IEEE Int. Conf. Image Process.* **1** 833–837.

THOMPSON, P. M., SCHWARTZ, C., LIN, R. T., KHAN, A. A. and TOGA, A. W. (1996). Three-dimensional statistical analysis of sulcal variability in the human brain. *J. Neurosci.* **16** 4261–4274.

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

VERCAUTEREN, T., PENNEC, X., PERCHANT, A. and AYACHE, N. (2009). Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage* **45** S61–S72. Mathematics in Brain Imaging.

VIERCK, C. J., WHITSEL, B. L., FAVOROV, O. V., BROWN, A. W. and TOMMERDAHL, M. (2013). Role of primary somatosensory cortex in the coding of pain. *Pain* **154** 334–344. https://doi.org/10.1016/j.pain.2012.10.021

VIOLA, P. and WELLS, W. M. (1997). Alignment by maximization of mutual information. *Int. J. Comput. Vis.* **24** 137–154.

VOGT, B. A., NIMCHINSKY, E. A., VOGT, L. J. and HOF, P. R. (1995). Human cingulate cortex: Surface features, flat maps, and cytoarchitecture. *J. Comp. Neurol.* **359** 490–506.

VOVK, V. G. (1990). Aggregating strategies. In Proc. of Computational Learning Theory. 1990.

WALKER, S. and HJORT, N. L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 811–821. MR1872068 https://doi.org/10.1111/1467-9868.00314

WANG, G., DATTA, A. and LINDQUIST, M. A. (2022). Supplement to "Bayesian functional registration of fMRI activation maps." https://doi.org/10.1214/21-AOAS1562SUPPA, https://doi.org/10.1214/21-AOAS1562SUPPB.

WATANABE, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11** 3571–3594. MR2756194

WOO, C.-W., ROY, M., BUHLE, J. T. and WAGER, T. D. (2015). Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. *PLoS Biol.* **13** e1002036. https://doi.org/10.1371/journal.pbio.1002036

XU, H., LORBERT, A., RAMADGE, P. J., GUNTUPALLI, J. S. and HAXBY, J. V. (2012). Regularized hyperalignment of multi-set fMRI data. In 2012 *IEEE Statistical Signal Processing Workshop*, *SSP* 2012 229–232.

YANG, C. and MEDIONI, G. (1992). Object modelling by registration of multiple range images. *Image Vis. Comput.* **10** 145–155.

YARKONI, T., POLDRACK, R. A., NICHOLS, T. E., VAN ESSEN, D. C. and WAGER, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8** 665–670.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1700–1717
https://doi.org/10.1214/21-AOAS1563

JOINT INTEGRATIVE ANALYSIS OF MULTIPLE DATA SOURCES WITH CORRELATED VECTOR OUTCOMES

BY EMILY C. HECTOR[1,a] AND PETER X.-K. SONG[2,b]

[1]*Department of Statistics, North Carolina State University,* [a]*ehector@ncsu.edu*
[2]*Department of Biostatistics, University of Michigan,* [b]*pxsong@umich.edu*

We propose a distributed quadratic inference function framework to jointly estimate regression parameters from multiple potentially heterogeneous data sources with correlated vector outcomes. The primary goal of this joint integrative analysis is to estimate covariate effects on all outcomes through a marginal regression model in a statistically and computationally efficient way. We develop a data integration procedure for statistical estimation and inference of regression parameters that is implemented in a fully distributed and parallelized computational scheme. To overcome computational and modeling challenges arising from the high-dimensional likelihood of the correlated vector outcomes, we propose to analyze each data source using Qu, Lindsay and Li's (*Biometrika* **87** (2000) 823–836) quadratic inference functions and then to jointly reestimate parameters from each data source by accounting for correlation between data sources using a combined meta-estimator in a similar spirit to the generalized method of moments put forward by Hansen (*Econometrica* **50** (1982) 1029–1054). We show both theoretically and numerically that the proposed method yields efficiency improvements and is computationally fast. We illustrate the proposed methodology with the joint integrative analysis of the association between smoking and metabolites in a large multicohort study and provide an R package for ease of implementation.

REFERENCES

ANDREWS, D. W. K. (1999). Consistent moment selection procedures for generalized method of moments estimation. *Econometrica* **67** 543–564. MR1685727 https://doi.org/10.1111/1468-0262.00036

CARAGEA, P. C. and SMITH, R. L. (2007). Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *J. Multivariate Anal.* **98** 1417–1440. MR2364128 https://doi.org/10.1016/j.jmva.2006.08.010

CHO, H. and QU, A. (2015). Efficient estimation for longitudinal data by combining large-dimensional moment conditions. *Electron. J. Stat.* **9** 1315–1334. MR3358326 https://doi.org/10.1214/15-EJS1036

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

DERSIMONIAN, R. and LAIRD, N. (2015). Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials* **45** 139–145.

FAN, J., HAN, F. and LIU, H. (2014). Challenges of big data analysis. *Nat. Sci. Rev.* **1** 293–314.

GLASS, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher* **5** 3–8.

GODAMBE, V. P. and HEYDE, C. C. (1987). Quasi-likelihood and optimal estimation. *Int. Stat. Rev.* **55** 231–244. MR0963141 https://doi.org/10.2307/1403403

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

HANSEN, L. P., HEATON, J. and YARON, A. (1996). Finite-sample properties of some alternative GMM estimators. *J. Bus. Econom. Statist.* **14** 262–280.

HECTOR, E. C. and SONG, P. X.-K. (2020). Doubly distributed supervised learning and inference with high-dimensional correlated outcomes. *J. Mach. Learn. Res.* **21** Paper No. 173. MR4209459

Key words and phrases. Data integration, generalized method of moments, parallel computing, quadratic inference function, scalable computing, seemingly unrelated regression.

HECTOR, E. C. and SONG, P. X.-K. (2022a). Supplement to "Joint integrative analysis of multiple data sources with correlated vector outcomes." https://doi.org/10.1214/21-AOAS1563SUPPA, https://doi.org/10.1214/21-AOAS1563SUPPB

HECTOR, E. C. and SONG, P. X.-K. (2022b). A distributed and integrated method of moments for high-dimensional correlated data analysis. *J. Amer. Statist. Assoc.* **116** 805–818. MR4270026 https://doi.org/10.1080/01621459.2020.1736082

HU, Y. and SONG, P. X.-K. (2012). Sample size determination for quadratic inference functions in longitudinal design with dichotomous outcomes. *Stat. Med.* **31** 787–800. MR2901800 https://doi.org/10.1002/sim.4458

IOANNIDIS, J. P. A. (2006). Meta-analysis in public health: Potentials and problems. *Italian Journal of Public Health* **3** 9–14.

JORDAN, M. I. (2013). On statistics, computation and scalability. *Bernoulli* **19** 1378–1390. MR3102908 https://doi.org/10.3150/12-BEJSP17

KUNDU, P., TANG, R. and CHATTERJEE, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika* **106** 567–585. MR3992390 https://doi.org/10.1093/biomet/asz030

LAAKSO, M., KUUSISTO, J., STANČÁKOVÁ, A., KUULASMAA, T., PAJUKANTA, P., LUSIS, A. J., COLLINS, F. S., MOHLKE, K. L. and BOEHNKE, M. (2017). The metabolic syndrome in men study: A resource for studies of metabolic and cardiovascular diseases. *J. Lipid. Res.* **58** 481–493. https://doi.org/10.1194/jlr.O072629

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

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

LIU, D., LIU, R. Y. and XIE, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: Efficiency and robustness. *J. Amer. Statist. Assoc.* **110** 326–340. MR3338506 https://doi.org/10.1080/01621459.2014.899235

NCBI (2021). PubChem Compound Summary for CID 1188, Xanthine. National Center for Biotechnology Information. Available at https://pubchem.ncbi.nlm.nih.gov/compound/Xanthine, Retrieved May 4, 2021.

QU, A., LINDSAY, B. G. and LI, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87** 823–836. MR1813977 https://doi.org/10.1093/biomet/87.4.823

SMITH, T. C., SPIEGELHALTER, D. J. and THOMAS, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Stat. Med.* **14** 2685–2699.

SONG, P. X.-K. (2007). *Correlated Data Analysis*: *Modeling*, *Analytics*, *and Applications*. *Springer Series in Statistics*. Springer, New York. MR2377853

SONG, P. X.-K., JIANG, Z., PARK, E. and QU, A. (2009). Quadratic inference functions in marginal models for longitudinal data. *Stat. Med.* **28** 3683–3696. MR2751731 https://doi.org/10.1002/sim.3719

TANG, L. and SONG, P. X. K. (2016). Fused lasso approach in regression coefficients clustering—learning parameter heterogeneity in data integration. *J. Mach. Learn. Res.* **17** Paper No. 113. MR3543519

TOULOUMIS, A. (2016). Simulating correlated binary and multinomial responses under marginal model specification: The SimCorMultRes package. *R J.* **8** 79–91.

VARIN, C. (2008). On composite marginal likelihoods. *AStA Adv. Stat. Anal.* **92** 1–28. MR2414624 https://doi.org/10.1007/s10182-008-0060-7

WANG, F., WANG, L. and SONG, P. X.-K. (2012). Quadratic inference function approach to merging longitudinal studies: Validation and joint estimation. *Biometrika* **99** 755–762. MR2966784 https://doi.org/10.1093/biomet/ass021

WANG, F., WANG, L. and SONG, P. X.-K. (2016). Fused lasso with the adaptation of parameter ordering in combining multiple studies with repeated measurements. *Biometrics* **72** 1184–1193. MR3591603 https://doi.org/10.1111/biom.12496

XIE, M. and SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *Int. Stat. Rev.* **81** 3–39. MR3047496 https://doi.org/10.1111/insr.12000

XIE, M., SINGH, K. and STRAWDERMAN, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *J. Amer. Statist. Assoc.* **106** 320–333. MR2816724 https://doi.org/10.1198/jasa.2011.tm09803

XU, T., HOLZAPFEL, C., DONG, X., BADER, E., YU, Z., PREHN, C., PERSTORFER, K., JAREMEK, M., ROEMISCH-MARGL, W. et al. (2013). Effects of smoking and smoking cessation on human serum metabolite profile: Results from the KORA cohort study. *BMC Med.* **11** 60.

YANG, G., LIU, D., LIU, R. Y., XIE, M. and HOAGLIN, D. C. (2014). Efficient network meta-analysis: A confidence distribution approach. *Stat. Methodol.* **20** 105–125. MR3205725 https://doi.org/10.1016/j.stamet.2014.01.003

ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57** 348–368. MR0139235

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1718–1746
https://doi.org/10.1214/21-AOAS1564

DETECTION OF TWO-WAY OUTLIERS IN MULTIVARIATE DATA AND APPLICATION TO CHEATING DETECTION IN EDUCATIONAL TESTS

BY YUNXIAO CHEN[a], YAN LU[b] AND IRINI MOUSTAKI[c]

Department of Statistics, London School of Economics and Political Science, [a]*y.chen186@lse.ac.uk,* [b]*y.lu62@lse.ac.uk,* [c]*i.moustaki@lse.ac.uk*

The paper proposes a new latent variable model for the simultaneous (two-way) detection of outlying individuals and items for item-response-type data. The proposed model is a synergy between a factor model for binary responses and continuous response times that captures normal item response behaviour and a latent class model that captures the outlying individuals and items. A statistical decision framework is developed under the proposed model that provides compound decision rules for controlling local false discovery/nondiscovery rates of outlier detection. Statistical inference is carried out under a Bayesian framework for which a Markov chain Monte Carlo algorithm is developed. The proposed method is applied to the detection of cheating in educational tests, due to item leakage, using a case study of a computer-based nonadaptive licensure assessment. The performance of the proposed method is evaluated by simulation studies.

REFERENCES

AGRESTI, A. and COULL, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *Amer. Statist.* **52** 119–126. MR1628435 https://doi.org/10.2307/2685469

ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. MR2549554 https://doi.org/10.1214/09-AOS689

ATCHADÉ, Y. F., ROBERTS, G. O. and ROSENTHAL, J. S. (2011). Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Stat. Comput.* **21** 555–568. MR2826692 https://doi.org/10.1007/s11222-010-9192-1

BAFUMI, J., GELMAN, A., PARK, D. K. and KAPLAN, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Polit. Anal.* **13** 171–187.

BARTHOLOMEW, D., KNOTT, M. and MOUSTAKI, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR2849614 https://doi.org/10.1002/9781119970583

BELOV, D. I. (2013). Detection of test collusion via Kullback–Leibler divergence. *J. Educ. Meas.* **50** 141–163.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BIRNBAUM, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores* (F. M. Lord and M. R. Novick, eds.) 397–472. Addison-Wesley, Oxford, England.

BOLT, D. M., COHEN, A. S. and WOLLACK, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *J. Educ. Meas.* **39** 331–348.

BOUGHTON, K. A. and YAMAMOTO, K. (2007). A HYBRID model for test speededness. In *Multivariate and Mixture Distribution Rasch Models* (M. von Davier and C. H. Carstensen, eds.) 147–156. Springer, New York, NY.

CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37 pp. MR2811000 https://doi.org/10.1145/1970392.1970395

CARLIN, B. P. and LOUIS, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. *Monographs on Statistics and Applied Probability* **69**. CRC Press, London. MR1427749

CASELLA, G. (1985). An introduction to empirical Bayes data analysis. *Amer. Statist.* **39** 83–87. MR0789118 https://doi.org/10.2307/2682801

Key words and phrases. Bayesian hierarchical model, outlier detection, false discovery rate, compound decision, test fairness, item response theory, latent class analysis.

CELEUX, G., HURN, M. and ROBERT, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95** 957–970. MR1804450 https://doi.org/10.2307/2669477

CHEN, Y., LEE, Y.-H. and LI, X. (2021). Item quality control in educational testing: Change point model, compound risk, and sequential detection. *J. Educ. Behav. Stat.* https://doi.org/10.3102/10769986211059085

CHEN, Y. and LI, X. (2020). Compound sequential change point detection in multiple data streams. *Statist. Sinica*. https://doi.org/10.5705/ss.202020.0508

CHEN, Y., LU, Y. and MOUSTAKI, I. (2022). Supplement to "Detection of two-way outliers in multivariate data and application to cheating detection in educational tests." https://doi.org/10.1214/21-AOAS1564SUPPA, https://doi.org/10.1214/21-AOAS1564SUPPB

CHO, S.-J., SUH, Y. and LEE, W.-Y. (2016). An NCME instructional module on latent DIF analysis using mixture item response models. *Educ. Meas., Issues Pract.* **35** 48–61.

CIZEK, G. J. and WOLLACK, J. A. (2017). *Handbook of Quantitative Methods for Detecting Cheating on Tests*. Routledge, New York, NY.

DOUGLAS, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika* **62** 7–28. MR1439472 https://doi.org/10.1007/BF02294778

DUNCAN, K. A. and MACEACHERN, S. N. (2008). Nonparametric Bayesian modelling for item response. *Stat. Model.* **8** 41–66. MR2750630 https://doi.org/10.1177/1471082X0700800104

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. MR2054289 https://doi.org/10.1198/016214504000000089

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866 https://doi.org/10.1214/07-STS236

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

EFRON, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statist. Sci.* **29** 285–301. MR3264543 https://doi.org/10.1214/13-STS455

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

EMBRETSON, S. E. and REISE, S. P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

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

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2** 1360–1383. MR2655663 https://doi.org/10.1214/08-AOAS191

GEYER, C. J. (2011). Importance sampling, simulated tempering, and umbrella sampling. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng, eds.). *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 295–311. CRC Press, Boca Raton, FL. MR2858453

GOEGEBEUR, Y., DE BOECK, P., WOLLACK, J. A. and COHEN, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika* **73** 65–87. MR2395293 https://doi.org/10.1007/s11336-007-9031-2

GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231. MR0370936 https://doi.org/10.1093/biomet/61.2.215

HADI, A. S. (1992). Identifying multiple outliers in multivariate data. *J. Roy. Statist. Soc. Ser. B* **54** 761–771. MR1185221

HOLLAND, P. W. and WAINER, H. (1993). *Differential Item Functioning*. Lawrence Erlbaum Associates, New York, NY.

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

KATZGRABER, H. G., TREBST, S., HUSE, D. A. and TROYER, M. (2006). Feedback-optimized parallel tempering Monte Carlo. *J. Stat. Mech. Theory Exp.* **2006** P03018.

KINGSTON, N. and CLARK, A. (2014). *Test Fraud*: *Statistical Detection and Methodology*. Routledge, New York, NY.

KUHA, J., KATSIKATSOU, M. and MOUSTAKI, I. (2018). Latent variable modelling with non-ignorable item non-response: Multigroup response propensity models for cross-national analysis. *J. Roy. Statist. Soc. Ser. A* **181** 1169–1192. MR3876387 https://doi.org/10.1111/rssa.12350

LAZARSFELD, P. F. and HENRY, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin Co., New York, NY.

LUNN, D., JACKSON, C., BEST, N., THOMAS, A. and SPIEGELHALTER, D. (2012). *The BUGS Book*: *A Practical Introduction to Bayesian Analysis*. CRC Press, Boca Raton, FL.

MAVRIDIS, D. and MOUSTAKI, I. (2008). Detecting outliers in factor analysis using the forward search algorithm. *Multivar. Behav. Res*. **43** 453–475.

MAVRIDIS, D. and MOUSTAKI, I. (2009). The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. *J. Comput. Graph. Statist*. **18** 1016–1034. MR2750449 https://doi.org/10.1198/jcgs.2009.08060

MCLEOD, L., LEWIS, C. and THISSEN, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Appl. Psychol. Meas*. **27** 121–137. MR1973750 https://doi.org/10.1177/0146621602250534

MILLSAP, R. E. (2012). *Statistical Approaches to Measurement Invariance*. Routledge, New York, NY.

MOUSTAKI, I. and VICTORIA-FESER, M.-P. (2006). Bounded-influence robust estimation in generalized linear latent variable models. *J. Amer. Statist. Assoc*. **101** 644–653. MR2256179 https://doi.org/10.1198/016214505000001320

O'LEARY, L. S. and SMITH, R. W. (2017). Detecting candidate preknowledge and compromised content using differential person and item functioning. In *Handbook of Quantitative Methods for Detecting Cheating on Tests* (G. J. Cizek and J. A. Wollack, eds.) 151–163. Routledge, New York, NY.

O'MUIRCHEARTAIGH, C. and MOUSTAKI, I. (1999). Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *J. Roy. Statist. Soc. Ser. A* **162** 177–194.

PISON, G., ROUSSEEUW, P. J., FILZMOSER, P. and CROUX, C. (2003). Robust factor analysis. *J. Multivariate Anal*. **84** 145–172. MR1965827 https://doi.org/10.1016/S0047-259X(02)00007-6

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

POOLE, K. T. and ROSENTHAL, H. (1991). Patterns of congressional voting. *Amer. J. Polit. Sci*. **35** 228–278.

POOLE, K. T., ROSENTHAL, H. and KOFORD, K. (1991). On dimensionalizing roll call votes in the US Congress. *Am. Polit. Sci. Rev*. **85** 955–976.

QUINTERO, A. and LESAFFRE, E. (2018). Comparing hierarchical models via the marginalized deviance information criterion. *Stat. Med*. **37** 2440–2454. MR3813295 https://doi.org/10.1002/sim.7649

RAMSAY, J. O. and WINSBERG, S. (1991). Maximum marginal likelihood estimation for semiparametric item analysis. *Psychometrika* **56** 365–379. MR1131766 https://doi.org/10.1007/BF02294480

RASCH, G. (1960). *Probabilistic Models for Some Intelligence and Achievement Tests*. Nielsen and Lydiche, Copenhagen, Denmark.

RECKASE, M. (2009). *Multidimensional Item Response Theory*. Springer, New York, NY.

REISER, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika* **61** 509–528.

RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792. MR1483213 https://doi.org/10.1111/1467-9868.00095

ROBBINS, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1950 (J. Neyman, ed.) 131–148. Univ. California Press, Berkeley, CA. MR0044803

ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* (J. Neyman, ed.) 157–163. Univ. California Press, Berkerly, CA. MR0084919

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

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist*. **6** 461–464. MR0468014

SEGALL, D. O. (2002). An item response model for characterizing test compromise. *J. Educ. Behav. Stat*. **27** 163–179.

SHAO, J. (2003). *Mathematical Statistics*: *Exercises and Solutions*. Springer, New York. MR2141827

SHU, Z., HENSON, R. and LUECHT, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika* **78** 481–497. MR3070232 https://doi.org/10.1007/s11336-012-9311-3

SINHARAY, S. (2017a). Detection of item preknowledge using likelihood ratio test and score test. *J. Educ. Behav. Stat*. **42** 46–68.

SINHARAY, S. (2017b). Which statistic should be used to detect item preknowledge when the set of compromised items is known? *Appl. Psychol. Meas*. **41** 403–421. https://doi.org/10.1177/0146621617698453

SKORUPSKI, W. P. and WAINER, H. (2017). The case for Bayesian methods when investigating test fraud. In *Handbook of Quantitative Methods for Detecting Cheating on Tests* (G. J. Cizek and J. A. Wollack, eds.) 214–231. Routledge, New York, NY.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. MR1979380 https://doi.org/10.1111/1467-9868.00353

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2014). The deviance information criterion: 12 years on. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 485–493. MR3210727 https://doi.org/10.1111/rssb.12062

SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657 https://doi.org/10.1198/016214507000000545

VAN DER LINDEN, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* **72** 287–308. MR2361958 https://doi.org/10.1007/s11336-006-1478-z

VEERKAMP, W. J. and GLAS, C. A. (2000). Detection of known items in adaptive testing with a statistical quality control method. *J. Educ. Behav. Stat.* **25** 373–389.

WANG, C., CHANG, H.-H. and DOUGLAS, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *Br. J. Math. Stat. Psychol.* **66** 144–168. MR3044864 https://doi.org/10.1111/j.2044-8317.2012.02045.x

WANG, X. and LIU, Y. (2020). Detecting compromised items using information from secure items. *J. Educ. Behav. Stat.* **45** 667–689.

WANG, C. and XU, G. (2015). A mixture hierarchical model for response times and response accuracy. *Br. J. Math. Stat. Psychol.* **68** 456–477.

WANG, C., XU, G. and SHANG, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika* **83** 223–254. MR3767020 https://doi.org/10.1007/s11336-016-9525-x

WIRTH, R. J. and EDWARDS, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychol. Methods* **12** 58–79. https://doi.org/10.1037/1082-989X.12.1.58

WOLLACK, J. A. and FREMER, J. J., eds. (2013). *Handbook of Test Security*. Routledge, New York, NY.

YUAN, K.-H. and BENTLER, P. M. (1998). Robust mean and covariance structure analysis. *Br. J. Math. Stat. Psychol.* **51** 63–88. MR1634841 https://doi.org/10.1111/j.2044-8317.1998.tb00667.x

YUAN, K.-H. and BENTLER, P. M. (2001). Effect of outliers on estimators and tests in covariance structure analysis. *Br. J. Math. Stat. Psychol.* **54** 161–175. MR1836858 https://doi.org/10.1348/000711001159366

ZHANG, C.-H. (2003). Compound decision theory and empirical Bayes methods. *Ann. Statist.* **31** 379–390. MR1983534 https://doi.org/10.1214/aos/1051027872

ZHANG, J. (2014). A sequential procedure for detecting compromised items in the item pool of a CAT system. *Appl. Psychol. Meas.* **38** 87–104.

ZHOU, Z., LI, X., WRIGHT, J., CANDÈS, E. and MA, Y. (2010). Stable principal component pursuit. In 2010 *IEEE International Symposium on Information Theory* 1518–1522. IEEE, New York.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1747–1773
https://doi.org/10.1214/21-AOAS1565

MEASUREMENT ERROR CORRECTION IN PARTICLE TRACKING MICRORHEOLOGY

BY YUN LING[1,a], MARTIN LYSY[1,b], IAN SEIM[2,c], JAY NEWBY[3,d], DAVID B. HILL[4,e], JEREMY CRIBB[5,f] AND M. GREGORY FOREST[6,g]

[1]*Department of Statistics and Actuarial Science, University of Waterloo,* [a]*y22ling@uwaterloo.ca,* [b]*mlysy@uwaterloo.ca*
[2]*Department of Applied Physical Sciences, University of North Carolina at Chapel Hill,* [c]*iseim@live.unc.edu*
[3]*Department of Mathematical and Statistical Sciences, University of Alberta,* [d]*jnewby@ualberta.ca*
[4]*Marsico Lung Institute, University of North Carolina at Chapel Hill,* [e]*dbhill@email.unc.edu*
[5]*Department of Physics and Astronomy, University of North Carolina at Chapel Hill,* [f]*jcribb@email.unc.edu*
[6]*Department of Biomedical Engineering, University of North Carolina at Chapel Hill,* [g]*forest@unc.edu*

In diverse biological applications, single-particle tracking (SPT) of passive microscopic species has become the experimental measurement of choice, when either the materials are of limited volume or so soft as to deform uncontrollably when manipulated by traditional instruments. In a wide range of SPT experiments, a ubiquitous finding is that of long-range dependence in the particles' motion. This is characterized by a power-law signature in the mean squared displacement (MSD) of particle positions as a function of time, the parameters of which reveal valuable information about the viscous and elastic properties of various biomaterials. However, MSD measurements are typically contaminated by complex and interacting sources of instrumental noise. As these often affect the high-frequency bandwidth to which MSD estimates are particularly sensitive, inadequate error correction can lead to severe bias in power law estimation and, thereby, the inferred viscoelastic properties. In this article we propose a novel strategy to filter high-frequency noise from SPT measurements. Our filters are shown theoretically to cover a broad spectrum of high-frequency noises and lead to a parametric estimator of MSD power-law coefficients for which an efficient computational implementation is presented. Based on numerous analyses of experimental and simulated data, results suggest our methods perform very well compared to other denoising procedures.

REFERENCES

AMBLARD, F., MAGGS, A. C., YURKE, B., PARGELLIS, A. N. and LEIBLER, S. (1996). Subdiffusion and anomalous local viscoelasticity in actin networks. *Phys. Rev. Lett.* **77** 4470.

AMMAR, G. S. and GRAGG, W. B. (1988). Superfast solution of real positive definite Toeplitz systems. *SIAM J. Matrix Anal. Appl.* **9** 61–76. MR0938136 https://doi.org/10.1137/0609005

ASHLEY, T. T. and ANDERSSON, S. B. (2015). Method for simultaneous localization and parameter estimation in particle tracking experiments. *Phys. Rev. E* **92** 052707.

BALCEREK, M., LOCH-OLSZEWSKA, H., TORRENO-PINA, J. A., GARCIA-PARAJO, M. F., WERON, A., MANZO, C. and BURNECKI, K. (2019). Inhomogeneous membrane receptor diffusion explained by a fractional heteroscedastic time series model. *Phys. Chem. Chem. Phys.* **21** 3114–3121. https://doi.org/10.1039/C8CP06781C

BERGLUND, A. J. (2010). Statistics of camera-based single-particle tracking. *Phys. Rev. E* **82** 011917.

BRIANE, V., KERVRANN, C. and VIMOND, M. (2018). Statistical analysis of particle trajectories in living cells. *Phys. Rev. E* **97** 062121.

BROCKWELL, P. J. and DAVIS, R. A. (1991). *Time Series*: *Theory and Methods*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1093459 https://doi.org/10.1007/978-1-4419-0320-4

BRONSTEIN, I., ISRAEL, Y., KEPTEN, E., MAI, S., SHAV-TAL, Y., BARKAI, E. and GARINI, Y. (2009). Transient anomalous diffusion of telomeres in the nucleus of mammalian cells. *Phys. Rev. Lett.* **103** 018102. https://doi.org/10.1103/PhysRevLett.103.018102

Key words and phrases. Single-particle tracking, subdiffusion, measurement error, high-frequency filtering.

BURNECKI, K., KEPTEN, E., GARINI, Y., SIKORA, G. and WERON, A. (2015). Estimating the anomalous diffusion exponent for single particle tracking data with measurement errors—an alternative approach. *Sci. Rep.* **5** 11306. https://doi.org/10.1038/srep11306

BURNECKI, K., SIKORA, G., WERON, A., TAMKUN, M. M. and KRAPF, D. (2019). Identifying diffusive motions in single-particle trajectories on the plasma membrane via fractional time-series models. *Phys. Rev. E* **99** 012101. https://doi.org/10.1103/PhysRevE.99.012101

BUROV, S., FIGLIOZZI, P., LIN, B., RICE, S. A., SCHERER, N. F. and DINNER, A. R. (2017). Single-pixel interior filling function approach for detecting and correcting errors in particle tracking. *Proc. Natl. Acad. Sci. USA* **114** 221–226.

CALDERON, C. P. (2016). Motion blur filtering: A statistical approach for extracting confinement forces and diffusivity from a single blurred trajectory. *Phys. Rev. E* **93** 053303. https://doi.org/10.1103/PhysRevE.93.053303

CHENOUARD, N., SMAL, I., DE CHAUMONT, F., MAŠKA, M., SBALZARINI, I. F., GONG, Y., CARDINALE, J., CARTHEL, C., CORALUPPI, S. et al. (2014). Objective comparison of particle tracking methods. *Nat. Methods* **11** 281–289.

CLAESKENS, G. and HJORT, N. L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.* **98** 900–945. With discussions and a rejoinder by the authors. MR2041482 https://doi.org/10.1198/016214503000000819

CRIBB, J., OSBORNE, L. D., VICCI, L., TAYLOR, R., HSIAO, H., O'BRIEN, E. T., HILL, D. B. and SUPERFINE, R. (2013). Panoptes: A 12 parallel microscope system for HCA. Developed at the Center for Computer Integrated Systems for Microscopy and Manipulation, University of North Carolina at Chapel Hill. Available at http://cismm.web.unc.edu/core-projects/force-microscopy/high-throughput-microscopy.

DESCHOUT, H., ZANACCHI, F. C., MLODZIANOSKI, M., DIASPRO, A., BEWERSDORF, J., HESS, S. T. and BRAECKMANS, K. (2014). Precisely and accurately localizing single emitters in fluorescence microscopy. *Nat. Methods* **11** 253–266.

DURBIN, J. (1960). The fitting of time-series models. *Rev. Inst. Int. Stat.* **28** 233–243. https://doi.org/10.2307/1401322

EDWARD, J. T. (1970). Molecular volumes and the Stokes–Einstein equation. *J. Chem. Educ.* **47** 261–270.

EINSTEIN, A. (1956). *Investigations on the Theory of the Brownian Movement*. Dover, New York. Edited with notes by R. Fürth, translated by A. D. Cowper. MR0077443

ERNST, M., JOHN, T., GUENTHER, M., WAGNER, C., SCHAEFER, U. F. and LEHR, C.-M. (2017). A model for the transient subdiffusive behavior of particles in mucus. *Biophys. J.* **112** 172–179.

FERRY, J. D. (1980). *Viscoelastic Properties of Polymers*. Wiley, New York, NY.

FONG, E. J., SHARMA, Y., FALLICA, B., TIERNEY, D. B., FORTUNE, S. M. and ZAMAN, M. H. (2013). Decoupling directed and passive motion in dynamic systems: Particle tracking microrheology of sputum. *Ann. Biomed. Eng.* **41** 837–846.

FREEDMAN, D. A. (2006). On the so-called "Huber sandwich estimator" and "robust standard errors". *Amer. Statist.* **60** 299–302. MR2291297 https://doi.org/10.1198/000313006X152207

GAL, N., LECHTMAN-GOLDSTEIN, D. and WEIHS, D. (2013). Particle tracking in living cells: A review of the mean square displacement method and beyond. *Rheol. Acta* **52** 425–443.

GEWEKE, J. and PORTER-HUDAK, S. (1983). The estimation and application of long memory time series models. *J. Time Series Anal.* **4** 221–238. MR0738585 https://doi.org/10.1111/j.1467-9892.1983.tb00371.x

GOULIAN, M. and SIMON, S. M. (2000). Tracking single proteins within cells. *Biophys. J.* **79** 2188–2198. https://doi.org/10.1016/S0006-3495(00)76467-8

GRØNNEBERG, S. and HJORT, N. L. (2014). The copula information criteria. *Scand. J. Stat.* **41** 436–459. MR3207180 https://doi.org/10.1111/sjos.12042

HANSEN, A. S., WORINGER, M., GRIMM, J. B., LAVIS, L. D., TJIAN, R. and DARZACQ, X. (2018). Robust model-based analysis of single-particle tracking experiments with spot-on. *eLife* **7**. https://doi.org/10.7554/eLife.33125

HERMANSEN, G. H., HJORT, N. L. and JULLUM, M. (2015). Parametric or nonparametric: The FIC approach for stationary time series. In *Proceedings of the* 60*th World Statistics Congress of the International Statistical Institute* 4827–4832. The International Statistical Institute.

HEYDE, C. C. (1997). *Quasi-Likelihood and Its Application*: *A General Approach to Optimal Parameter Estimation*. *Springer Series in Statistics*. Springer, New York. MR1461808 https://doi.org/10.1007/b98823

HILL, D. B., VASQUEZ, P. A., MELLNIK, J., MCKINLEY, S. A., VOSE, A., MU, F., HENDERSON, A. G., DONALDSON, S. H., ALEXIS, N. E. et al. (2014). A biophysical basis for mucus solids concentration as a candidate biomarker for airways disease. *PLoS ONE* **9** e87681.

KAILATH, T., KUNG, S. Y. and MORF, M. (1979). Displacement ranks of matrices and linear equations. *J. Math. Anal. Appl.* **68** 395–407. MR0533501 https://doi.org/10.1016/0022-247X(79)90124-0

KOSLOVER, E. F., CHAN, C. K. and THERIOT, J. A. (2016). Disentangling random motion and flow in a complex medium. *Biophys. J.* **110** 700–709. https://doi.org/10.1016/j.bpj.2015.11.008

KOU, S. C. (2008). Stochastic modeling in nanoscale biophysics: Subdiffusion within proteins. *Ann. Appl. Stat.* **2** 501–535. MR2524344 https://doi.org/10.1214/07-AOAS149

KOWALCZYK, A., OELSCHLAEGER, C. and WILLENBACHER, N. (2014). Tracking errors in 2D multiple particle tracking microrheology. *Meas. Sci. Technol.* **26** 015302.

KOWALEK, P., LOCH-OLSZEWSKA, H. and SZWABIŃSKI, J. (2019). Classification of diffusion modes in single-particle tracking data: Feature-based versus deep-learning approach. *Phys. Rev. E* **100** 032410. https://doi.org/10.1103/PhysRevE.100.032410

KUBO, R. (1966). The fluctuation-dissipation theorem. *Rep. Progr. Phys.* **29** 255–284.

LAI, S. K., O'HANLON, D. E., HARROLD, S., MAN, S. T., WANG, Y.-Y., CONE, R. and HANES, J. (2007). Rapid transport of large polymeric nanoparticles in fresh undiluted human mucus. *Proc. Natl. Acad. Sci. USA* **104** 1482–1487.

LEE, S.-H., ROICHMAN, Y., YI, G.-R., KIM, S.-H., YANG, S.-M., VAN BLAADEREN, A., VAN OOSTRUM, P. and GRIER, D. G. (2007). Characterizing and tracking single colloidal particles with video holographic microscopy. *Opt. Express* **15** 18275–18282.

LEVINSON, N. (1947). The Wiener RMS (root mean square) error criterion in filter design and prediction. *J. Math. Phys.* **25** 261–278. MR0019257 https://doi.org/10.1002/sapm1946251261

LING, Y. and LYSY, M. (2017). `SuperGauss`: Superfast likelihood inference for stationary Gaussian time series. R package version 2.0.2. Available at https://CRAN.R-project.org/package=SuperGauss.

LING, Y., LYSY, M., SEIM, I., NEWBY, J., HILL, D. B, CRIBB, J. and FOREST, M. G., (2022). Supplement to "Measurement error correction in particle tracking microrheology." https://doi.org/10.1214/21-AOAS1565SUPPA, https://doi.org/10.1214/21-AOAS1565SUPPB

LYSY, M., PILLAI, N. S., HILL, D. B., FOREST, M. G., MELLNIK, J. W. R., VASQUEZ, P. A. and MCKINLEY, S. A. (2016). Model comparison and assessment for single particle tracking in biological fluids. *J. Amer. Statist. Assoc.* **111** 1413–1426. MR3601698 https://doi.org/10.1080/01621459.2016.1158716

LYSY, M. and LING, Y. (2021). `subdiff`: Subdiffusive modeling in passive particle-tracking microrheology. R package version 0.0.1. Available at https://github.com/mlysy/subdiff.

MASON, T. G. and WEITZ, D. A. (1995). Optical measurements of frequency-dependent linear viscoelastic moduli of complex fluids. *Phys. Rev. Lett.* **74** 1250–1253. https://doi.org/10.1103/PhysRevLett.74.1250

MASON, T. G., GANESAN, K., VAN ZANTEN, J. H., WIRTZ, D. and KUO, S. C. (1997). Particle tracking microrheology of complex fluids. *Phys. Rev. Lett.* **79** 3282–3285. https://doi.org/10.1103/PhysRevLett.79.3282

MAZZA, D., ABERNATHY, A., GOLOB, N., MORISAKI, T. and MCNALLY, J. G. (2012). A benchmark for chromatin binding measurements in live cells. *Nucleic Acids Res.* **40** e119. https://doi.org/10.1093/nar/gks701

MCKINLEY, S. A., YAO, L. and FOREST, M. G. (2009). Transient anomalous diffusion of tracer particles in soft matter. *J. Rheol.* **53** 1487–1506.

MELLNIK, J. W., LYSY, M., VASQUEZ, P. A., PILLAI, N. S., HILL, D. B., CRIBB, J., MCKINLEY, S. A. and FOREST, M. G. (2016). Maximum likelihood estimation for single particle, passive microrheology data with drift. *J. Rheol.* **60** 379–392.

MICHALET, X. (2010). Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium. *Phys. Rev. E* **82** 041914, 13. MR2788037 https://doi.org/10.1103/PhysRevE.82.041914

MICHALET, X. and BERGLUND, A. J. (2012). Optimal diffusion coefficient estimation in single-particle tracking. *Phys. Rev. E* **85** 061916.

MONNIER, N., GUO, S. M., MORI, M., HE, J., LÉNÁRT, P. and BATHE, M. (2012). Bayesian approach to MSD-based analysis of particle motion in live cells. *Biophys. J.* **103** 616–626.

MONNIER, N., BARRY, Z., PARK, H. Y., SU, K.-C., KATZ, Z., ENGLISH, B. P., DEY, A., PAN, K., CHEESEMAN, I. M. et al. (2015). Inferring transient particle transport dynamics in live cells. *Nat. Methods* **12** 838–840. https://doi.org/10.1038/nmeth.3483

MORRIS, J. S. (2015). Functional regression. *Annu. Rev. Stat. Appl.* **2** 321–359. https://doi.org/10.1146/annurev-statistics-010814-020413

MORTENSEN, K. I., CHURCHMAN, L. S., SPUDICH, J. A. and FLYVBJERG, H. (2010). Optimized localization analysis for single-molecule tracking and super-resolution microscopy. *Nat. Methods* **7** 377–381.

NEWBY, J. M., SCHAEFER, A. M., LEE, P. T., FOREST, M. G. and LAI, S. K. (2018). Convolutional neural networks automate detection for tracking of submicron-scale particles in 2D and 3D. *Proc. Natl. Acad. Sci. USA* **115** 9026–9031.

PERSSON, F., LINDÉN, M., UNOSON, C. and ELF, J. (2013). Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nat. Methods* **10** 265–269. https://doi.org/10.1038/nmeth.2367

QIAN, H., SHEETZ, M. P. and ELSON, E. L. (1991). Single particle tracking. Analysis of diffusion and flow in two-dimensional systems. *Biophys. J.* **60** 910–921.

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

ROWLANDS, C. J. and SO, P. T. (2013). On the correction of errors in some multiple particle tracking experiments. *Appl. Phys. Lett.* **102** 021913.

SAVIN, T. and DOYLE, P. S. (2005). Static and dynamic errors in particle tracking microrheology. *Biophys. J.* **88** 623–638.

SAXTON, M. J. and JACOBSON, K. (1997). Single-particle tracking: Applications to membrane dynamics. *Annu. Rev. Biophys. Biomol. Struct.* **26** 373–399.

SIKORA, G., TEUERLE, M., WYŁOMAŃSKA, A. and GREBENKOV, D. (2017a). Statistical properties of the anomalous scaling exponent estimator based on time-averaged mean-square displacement. *Phys. Rev. E* **96** 022132. https://doi.org/10.1103/PhysRevE.96.022132

SIKORA, G., KEPTEN, E., WERON, A., BALCEREK, M. and BURNECKI, K. (2017b). An efficient algorithm for extracting the magnitude of the measurement error for fractional dynamics. *Phys. Chem. Chem. Phys.* **19** 26566–26581. https://doi.org/10.1039/C7CP04464J

SOUSSOU, J. E., MOAVENZADEH, F. and GRADOWCZYK, M. H. (1970). Application of Prony series to linear viscoelasticity. *Trans. Soc. Rheol.* **14** 573–584.

SUH, J., DAWSON, M. and HANES, J. (2005). Real-time multiple-particle tracking: Applications to drug and gene delivery. *Adv. Drug Deliv. Rev.* **57** 63–78. https://doi.org/10.1016/j.addr.2004.06.001

SZYMANSKI, J. and WEISS, M. (2009). Elucidating the origin of anomalous diffusion in crowded fluids. *Phys. Rev. Lett.* **103** 038102. https://doi.org/10.1103/PhysRevLett.103.038102

TAYLOR, R., HSIAO, J., HAHN, P. and CRIBB, J. (2018). Video spot tracker. Developed at the Center for Computer Integrated Systems for Microscopy and Manipulation, University of North Carolina at Chapel Hill. Available at http://cismm.web.unc.edu/resources/software-manuals/video-spot-tracker-manual.

TELEDYNE FLIR (2019). Flea3 USB3 camera. Available at https://www.flir.com/products/flea3-usb3/.

TÜRKCAN, S. and MASSON, J.-B. (2013). Bayesian decision tree for the classification of the mode of motion in single-molecule trajectories. *PLoS ONE* **8** e82799. https://doi.org/10.1371/journal.pone.0082799

VAN DER SCHAAR, H. M., RUST, M. J., CHEN, C., VAN DER ENDE-METSELAAR, H., WILSCHUT, J., ZHUANG, X. and SMIT, J. M. (2008). Dissecting the cell entry pathway of Dengue virus by single-particle tracking in living cells. *PLoS Pathog.* **4** e1000244.

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

VESTERGAARD, C. L., BLAINEY, P. C. and FLYVBJERG, H. (2014). Optimal estimation of diffusion coefficients from single-particle trajectories. *Phys. Rev. E* **89** 022726.

WANG, Y.-Y., LAI, S. K., SUK, J. S., PACE, A., CONE, R. and HANES, J. (2008). Addressing the PEG mucoadhesivity paradox to engineer nanoparticles that "slip" through the human mucus barrier. *Angew. Chem.* **47** 9726–9729.

WEIHS, D., TEITELL, M. A. and MASON, T. G. (2007). Simulations of complex particle transport in heterogeneous active liquids. *Microfluid. Nanofluid.* **3** 227–237.

WEISS, M. (2013). Single-particle tracking data reveal anticorrelated fractional Brownian motion in crowded fluids. *Phys. Rev. E* **88** 010101.

WEISS, M., ELSNER, M., KARTBERG, F. and NILSSON, T. (2004). Anomalous subdiffusion is a measure for cytoplasmic crowding in living cells. *Biophys. J.* **87** 3518–3524.

WIRTZ, D. (2009). Particle-tracking microrheology of living cells: Principles and applications. *Annu. Rev. Biophys.* **38** 301–326.

WONG, I. Y., GARDEL, M. L., REICHMAN, D. R., WEEKS, E. R., VALENTINE, M. T., BAUSCH, A. R. and WEITZ, D. A. (2004). Anomalous diffusion probes microstructure dynamics of entangled F-actin networks. *Phys. Rev. Lett.* **92** 178101. https://doi.org/10.1103/PhysRevLett.92.178101

WORKING, P. K., NEWMAN, M. S., JOHNSON, J. and CORNACOFF, J. B. (1997). Safety of poly(ethylene glycol) and poly(ethylene glycol) derivatives. In *Poly*(*Ethylene Glycol*). *ACS Symposium Series* **680** 45–57 4. Am. Chem. Soc., Washington. https://doi.org/10.1021/bk-1997-0680.ch004

ZHANG, K., CRIZER, K. P. R., SCHOENFISCH, M. H., HILL, D. B. and DIDIER, G. (2018). Fluid heterogeneity detection based on the asymptotic distribution of the time-averaged mean squared displacement in single particle tracking experiments. *J. Phys. A* **51** 445601, 41. MR3863293 https://doi.org/10.1088/1751-8121/aae0af

ZWANZIG, R. (2001). *Nonequilibrium Statistical Mechanics*. Oxford Univ. Press, New York. MR2012558

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1774–1794
https://doi.org/10.1214/21-AOAS1566

SENSITIVITY ANALYSIS FOR EVALUATING PRINCIPAL SURROGATE ENDPOINTS RELAXING THE EQUAL EARLY CLINICAL RISK ASSUMPTION

BY YING HUANG[a], YINGYING ZHUANG[b] AND PETER GILBERT[c]

Fred Hutchinson Cancer Research Center, [a]*yhuang@fredhutch.org,* [b]*yyzhuang@uw.edu,* [c]*pgilbert@fredhutch.org*

This article addresses the evaluation of postrandomization immune response biomarkers as principal surrogate endpoints of a vaccine's protective effect, based on data from randomized vaccine trials. An important metric for quantifying a biomarker's principal surrogacy in vaccine research is the vaccine efficacy curve, which shows a vaccine's efficacy as a function of potential biomarker values if receiving vaccine, among an "early-always-at-risk" principal stratum of trial participants who remain disease-free at the time of biomarker measurement whether having received vaccine or placebo. Earlier work in principal surrogate evaluation relied on an "equal-early-clinical-risk" assumption for identifiability of the vaccine curve, based on observed disease status at the time of biomarker measurement. This assumption is violated in the common setting that the vaccine has an early effect on the clinical endpoint before the biomarker is measured. In particular, a vaccine's early protective effect observed in two phase III dengue vaccine trials (CYD14/CYD15) has motivated our current research development. We relax the "equal-early-clinical-risk" assumption and propose a new sensitivity analysis framework for principal surrogate evaluation allowing for early vaccine efficacy. Under this framework we develop inference procedures for vaccine efficacy curve estimators, based on the estimated maximum likelihood approach. We then use the proposed methodology to assess the surrogacy of postrandomization neutralization titer in the motivating dengue application.

REFERENCES

BADEN, L. R., EL SAHLY, H. M., ESSINK, B., KOTLOFF, K., FREY, S., NOVAK, R., DIEMERT, D., SPECTOR, S. A., ROUPHAEL, N. et al. (2021). Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N. Engl. J. Med.* **384** 403–416.

BURZYKOWSKI, T., MOLENBERGHS, G. and BUYSE, M. (2006). *The Evaluation of Surrogate Endpoints*. Springer Science & Business Media.

BUYSE, M., MOLENBERGHS, G., BURZYKOWSKI, T., RENARD, D. and GEYS, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* **1** 49–67. https://doi.org/10.1093/biostatistics/1.1.49

CAPEDING, M. R., TRAN, N. H., HADINEGORO, S. R. S., ISMAIL, H. I. H. M., CHOTPITAYASUNONDH, T., CHUA, M. N. et al. (2014). Clinical efficacy and safety of a novel tetravalent Dengue vaccine in healthy children in Asia: A phase 3, randomised, observer-masked, placebo-controlled trial. *Lancet* **384** 1358–1365.

DANIELS, M. J. and HUGHES, M. D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Stat. Med.* **16** 1965–1982.

FOLLMANN, D. (2006). Augmented designs to assess immune response in vaccine trials. *Biometrics* **62** 1161–1169. MR2307441 https://doi.org/10.1111/j.1541-0420.2006.00569.x

FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039 https://doi.org/10.1111/j.0006-341X.2002.00021.x

FREEDMAN, L. S., GRAUBARD, B. I. and SCHATZKIN, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Stat. Med.* **11** 167–178.

GABRIEL, E. E. and GILBERT, P. B. (2014). Evaluating principal surrogate endpoints with time-to-event data accounting for time-varying treatment efficacy. *Biostatistics* **15** 251–265. https://doi.org/10.1093/biostatistics/kxt055

Key words and phrases. Causal inference, estimated maximum likelihood, principal stratification, principal surrogate, sensitivity analysis, vaccine efficacy curve.

GILBERT, P. B., BOSCH, R. J. and HUDGENS, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59** 531–541. MR2004258 https://doi.org/10.1111/1541-0420.00063

GILBERT, P. B. and HUDGENS, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64** 1146–1154. MR2522262 https://doi.org/10.1111/j.1541-0420.2008.01014.x

GILBERT, P. B., BLETTE, B. S., SHEPHERD, B. E. and HUDGENS, M. G. (2020). Post-randomization biomarker effect modification analysis in an HIV vaccine clinical trial. *J. Causal Inference* **8** 54–69.

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

HUANG, Y. (2018). Evaluating principal surrogate markers in vaccine trials in the presence of multiphase sampling. *Biometrics* **74** 27–39. MR3777923 https://doi.org/10.1111/biom.12737

HUANG, Y., GILBERT, P. B. and WOLFSON, J. (2013). Design and estimation for evaluating principal surrogate markers in vaccine trials. *Biometrics* **69** 301–309. MR3071048 https://doi.org/10.1111/biom.12014

HUANG, Y., ZHUANG, Y. and GILBERT, P. (2022). Supplement to "Sensitivity analysis for evaluating principal surrogate endpoints relaxing the equal early clinical risk assumption." https://doi.org/10.1214/21-AOAS1566SUPP

JOFFE, M. M. and GREENE, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65** 530–538. MR2751477 https://doi.org/10.1111/j.1541-0420.2008.01106.x

LI, Y., TAYLOR, J. M., ELLIOTT, M. R. and SARGENT, D. J. (2011). Causal assessment of surrogacy in a meta-analysis of colorectal cancer trials. *Biostatistics* **12** 478–492.

LIN, D., FLEMING, T. and DE GRUTTOLA, V. (1997). Estimating the proportion of treatment effect explained by a surrogate marker. *Stat. Med.* **16** 1515–1527.

MOLENBERGHS, G., KENWARD, M. G. and GOETGHEBEUR, E. (2001). Sensitivity analysis for incomplete contingency tables: The Slovenian plebiscite case. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **50** 15–29.

MOODIE, Z., JURASKA, M., HUANG, Y., ZHUANG, Y., FONG, Y., CARPP, L. et al. (2018). Neutralizing antibody correlates analysis of tetravalent Dengue vaccine efficacy trials in Asia and Latin America. *J. Infect. Dis.* **217** 742–753.

PEPE, M. S. and FLEMING, T. R. (1991). A nonparametric method for dealing with mismeasured covariate data. *J. Amer. Statist. Assoc.* **86** 108–113. MR1137103

PRENTICE, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat. Med.* **8** 431–440.

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

SHEPHERD, B. E., GILBERT, P. B., JEMIAI, Y. and ROTNITZKY, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics* **62** 332–342. MR2236845 https://doi.org/10.1111/j.1541-0420.2005.00495.x

VANSTEELANDT, S., GOETGHEBEUR, E., KENWARD, M. G. and MOLENBERGHS, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statist. Sinica* **16** 953–979. MR2281311

VILLAR, L., DAYAN, G. H., ARREDONDO-GARCÍA, J. L., RIVERA, D. M., CUNHA, R., DESEDA, C. et al. (2015). Efficacy of a tetravalent Dengue vaccine in children in Latin America. *N. Engl. J. Med.* **372** 113–123.

WOLFSON, J. and GILBERT, P. (2010). Statistical identifiability and the surrogate endpoint problem, with application to vaccine trials. *Biometrics* **66** 1153–1161. MR2758503 https://doi.org/10.1111/j.1541-0420.2009.01380.x

ZHUANG, Y., HUANG, Y. and GILBERT, P. B. (2019). Simultaneous inference of treatment effect modification by intermediate response endpoint principal strata with application to vaccine trials. *Int. J. Biostat.*

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1795–1821
https://doi.org/10.1214/21-AOAS1567

PARAMETER CALIBRATION IN WAKE EFFECT SIMULATION MODEL WITH STOCHASTIC GRADIENT DESCENT AND STRATIFIED SAMPLING

BY BINGJIE LIU[1,a], XUBO YUE[2,b], EUNSHIN BYON[2,c] AND RAED AL KONTAR[2,d]

[1]*Zhejiang Lab,* [a]*bingjiel@zhejianglab.edu*

[2]*Department of Industrial and Operations Engineering, University of Michigan,* [b]*maxyxb@umich.edu,* [c]*ebyon@umich.edu,* [d]*alkontar@umich.edu*

As the market share of wind energy has been rapidly growing, wake effect analysis is gaining substantial attention in the wind industry. Wake effects represent a wind shade cast by upstream turbines to the downwind direction, resulting in power deficits in downstream turbines. To quantify the aggregated influence of wake effects on the power generation of a wind farm, various simulation models have been developed, including Jensen's wake model. These models include parameters that need to be calibrated from field data. Existing calibration methods are based on surrogate models that impute the data under the assumption that physical and/or computer trials are computationally expensive, typically at the design stage. This, however, is not the case where large volumes of data can be collected during the operational stage. Motivated by the wind energy application, we develop a new calibration approach for big data settings without the need for statistical emulators. Specifically, we cast the problem into a stochastic optimization framework and employ stochastic gradient descent to iteratively refine calibration parameters using randomly selected subsets of data. We then propose a stratified sampling scheme that enables choosing more samples from noisy and influential sampling regions and thus reducing the variance of the estimated gradient for improved convergence. Through both theoretical and numerical studies on wind farm data, we highlight the benefits of our variance-conscious calibration approach.

REFERENCES

AINSLIE, J. F. (1988). Calculating the flowfield in the wake of wind turbines. *J. Wind Eng. Ind. Aerodyn.* **27** 213–224.

ÁLVAREZ, M. A. and LAWRENCE, N. D. (2011). Computationally efficient convolved multiple output Gaussian processes. *J. Mach. Learn. Res.* **12** 1459–1500. MR2813145

BARTHELMIE, R. J. and PRYOR, S. (2013). An overview of data for wake model evaluation in the Virtual Wakes Laboratory. *Appl. Energy* **104** 834–844.

BARTHELMIE, R. J., PRYOR, S. C., FRANDSEN, S. T., HANSEN, K. S., SCHEPERS, J. G., RADOS, K., SCHLEZ, W., NEUBERT, A., JENSEN, L. E. et al. (2010). Quantifying the impact of wind turbine wakes on power output at offshore wind farms. *Journal of Atmospheric and Oceanic Technology* **27** 1302–1317.

BOLLAPRAGADA, R., BYRD, R. and NOCEDAL, J. (2018). Adaptive sampling strategies for stochastic optimization. *SIAM J. Optim.* **28** 3312–3343. MR3890789 https://doi.org/10.1137/17M1154679

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees. Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

CHEN, X., LEE, J. D., TONG, X. T. and ZHANG, Y. (2020). Statistical inference for model parameters in stochastic gradient descent. *Ann. Statist.* **48** 251–273. MR4065161 https://doi.org/10.1214/18-AOS1801

CHOE, Y., BYON, E. and CHEN, N. (2015). Importance sampling for reliability evaluation with stochastic simulation models. *Technometrics* **57** 351–361. MR3384950 https://doi.org/10.1080/00401706.2014.1001523

CHOE, Y., PAN, Q. and BYON, E. (2016). Computationally efficient uncertainty minimization in wind turbine extreme load assessments. *J. Sol. Energy Eng.* **138** 041012.

CHURCHFIELD, M. (2013). Review of Wind Turbine Wake Models and Future Directions (Presentation). Technical report, National Renewable Energy Laboratory (NREL), Golden, CO.

Key words and phrases. Jensen's wake model, stochastic optimization, variance reduction, wind energy.

DAMIANOU, A. C., TITSIAS, M. K. and LAWRENCE, N. D. (2016). Variational inference for latent variables and uncertain inputs in Gaussian processes. *J. Mach. Learn. Res.* **17** Paper No. 42. MR3491136

DTU WIND ENERGY (2015). Wind resources for energy production of wind turbines. Available at http://www.wasp.dk/wasp#details__wakeeffectmodel Accessed: 2015-09-18.

FANG, Y., XU, J. and YANG, L. (2018). Online bootstrap confidence intervals for the stochastic gradient descent estimator. *J. Mach. Learn. Res.* **19** Paper No. 78. MR3899780

FRANDSEN, S. (1992). On the wind speed reduction in the center of large clusters of wind turbines. *J. Wind Eng. Ind. Aerodyn.* **39** 251–265.

GÖÇMEN, T., VAN DER LAAN, P., RÉTHORÉ, P.-E., DIAZ, A. P., LARSEN, G. C. and OTT, S. (2016). Wind turbine wake models developed at the technical university of Denmark: A review. *Renew. Sustain. Energy Rev.* **60** 752–769.

GRAMACY, R. B., BINGHAM, D., HOLLOWAY, J. P., GROSSKOPF, M. J., KURANZ, C. C., RUTTER, E., TRANTHAM, M. and DRAKE, P. R. (2015). Calibrating a large computer experiment simulating radiative shock hydrodynamics. *Ann. Appl. Stat.* **9** 1141–1168. MR3418718 https://doi.org/10.1214/15-AOAS850

HANKIN, R. (2019). calibrator: Bayesian calibration of complex computer codes. R Package, Version 1.2-8.

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

INTERNATIONAL ENERGY AGENCY (2015). Medium-Term Renewable Energy Market Report 2015—Market Analysis and Forecasts to 2020. Technical report.

JENSEN, N. O. (1983). *A Note on Wind Generator Interaction*.

JOSEPH, V. R. and YAN, H. (2015). Engineering-driven statistical adjustment and calibration. *Technometrics* **57** 257–267. MR3369681 https://doi.org/10.1080/00401706.2014.902773

KATIC, I., HØJSTRUP, J. and JENSEN, N. (1986). A simple model for cluster efficiency. In *European Wind Energy Association Conference and Exhibition* 407–410.

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

KHEIRABADI, A. C. and NAGAMUNE, R. (2019). A quantitative review of wind farm control with the objective of wind farm power maximization. *J. Wind Eng. Ind. Aerodyn.* **192** 45–73.

LARSEN, G. C. (1988). A simple wake calculation procedure. Technical report, Risø National Laboratory, Denmark.

LEE, G., BYON, E., NTAIMO, L. and DING, Y. (2013). Bayesian spline method for assessing extreme loads on wind turbines. *Ann. Appl. Stat.* **7** 2034–2061. MR3161712 https://doi.org/10.1214/13-AOAS670

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

LIU, B., YUE, X., BYON, E. and KONTAR, R. A (2022). Supplement to "Parameter calibration in wake effect simulation model with stochastic gradient descent and stratified sampling." https://doi.org/10.1214/21-AOAS1567SUPP

LOHR, S. L. (2010). *Sampling*: *Design and Analysis*, 2nd ed. Brooks/Cole, Cengage Learning, Boston, MA. MR3057878

MEASE, D. and BINGHAM, D. (2006). Latin hyperrectangle sampling for computer experiments. *Technometrics* **48** 467–477. MR2328616 https://doi.org/10.1198/004017006000000101

MOLLAND, A. F. and TURNOCK, S. R. (2011). *Marine Rudders and Control Surfaces*: *Principles*, *Data*, *Design and Applications*. Elsevier, Amsterdam.

NING, S., BYON, E., WU, T. and LI, J. (2017). A sparse partitioned-regression model for nonlinear system–environment interactions. *IISE Trans.* **49** 814–826.

OWEN, A. B. (2013). Monte Carlo theory, methods and examples. Book in progress. Online version available at https://statweb.stanford.edu/~owen/mc/.

PAQUETTE, C. and SCHEINBERG, K. (2020). A stochastic line search method with expected complexity analysis. *SIAM J. Optim.* **30** 349–376. MR4060460 https://doi.org/10.1137/18M1216250

PAULO, R., GARCÍA-DONATO, G. and PALOMO, J. (2012). Calibration of computer models with multivariate output. *Comput. Statist. Data Anal.* **56** 3959–3974. MR2957846 https://doi.org/10.1016/j.csda.2012.05.023

POLYAK, B. T. (1990). New stochastic approximation type procedures. *Autom. Remote Control* **7** 937–1008.

POLYAK, B. T. and JUDITSKY, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30** 838–855. MR1167814 https://doi.org/10.1137/0330046

QIAN, P. Z. G. and WU, C. F. J. (2008). Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics* **50** 192–204. MR2439878 https://doi.org/10.1198/004017008000000082

RASMUSSEN, C. E. (2003). Gaussian processes in machine learning. In *Summer School on Machine Learning* 63–71. Springer, Berlin.

RUPPERT, D. (1988). Efficient estimations from a slowly convergent Robbins–Monro process. Technical report, Cornell Univ. Operations Research and Industrial Engineering.

SHASHAANI, S., HASHEMI, F. S. and PASUPATHY, R. (2018). ASTRO-DF: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. *SIAM J. Optim.* **28** 3145–3176. MR3880261 https://doi.org/10.1137/15M1042425

SNELSON, E. and GHAHRAMANI, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems* 1257–1264.

STAID, A. (2015). Statistical modeling to support power system planning. Ph.D. thesis, Johns Hopkis Univ., Washington, DC.

THERNEAU, T. M. and ATKINSON, E. J. (2019). An introduction to recursive partitioning using the RPART routines. R Package.

THERNEAU, T., ATKINSON, B., RIPLEY, B. and RIPLEY, M. B. (2019). rpart: Recursive partitioning and regression trees. R Package, Version 4.1-15.

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

TUO, R. and WU, C. F. J. (2016). A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties. *SIAM/ASA J. Uncertain. Quantificat.* **4** 767–795. MR3523087 https://doi.org/10.1137/151005841

YOU, M., BYON, E., JIN, J. J. and LEE, G. (2017). When wind travels through turbines: A new statistical approach for characterizing heterogeneous wake effects in multi-turbine wind farms. *IISE Trans.* **49** 84–95.

YOU, M., LIU, B., BYON, E., HUANG, S. and JIN, J. (2018). Direction-dependent power curve modeling for multiple interacting wind turbines. *IEEE Trans. Power Syst.* **33** 1725–1733.

YUAN, J., NG, S. H. and TSUI, K. L. (2013). Calibration of stochastic computer models using stochastic approximation methods. *IEEE Trans. Autom. Sci. Eng.* **10** 171–186.

ZWAKMAN, J. (2014). Wind turbines and the environment. Available at https://www.wijkplatformsvelsen.nl/ijmuiden-noord/2014/11/23/windturbines-en-het-milieu/ Accessed: 2020-09-11.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1822–1847
https://doi.org/10.1214/21-AOAS1568

ASYMMETRIC TAIL DEPENDENCE MODELING, WITH APPLICATION TO CRYPTOCURRENCY MARKET DATA

BY YAN GONG[a] AND RAPHAËL HUSER[b]

Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), [a]*yan.gong@kaust.edu.sa,* [b]*raphael.huser@kaust.edu.sa*

Since the inception of Bitcoin in 2008, cryptocurrencies have played an increasing role in the world of e-commerce, but the recent turbulence in the cryptocurrency market in 2018 has raised some concerns about their stability and associated risks. For investors it is crucial to uncover the dependence relationships between cryptocurrencies for a more resilient portfolio diversification. Moreover, the stochastic behavior in both tails is important, as long positions are sensitive to a decrease in prices (lower tail), while short positions are sensitive to an increase in prices (upper tail). In order to assess both risk types, we develop in this paper a flexible copula model which is able to distinctively capture asymptotic dependence or independence in its lower and upper tails simultaneously. Our proposed model is parsimonious and smoothly bridges (in each tail) both extremal dependence classes in the interior of the parameter space. Inference is performed using a full or censored likelihood approach, and we investigate by simulation the estimators' efficiency under three different censoring schemes which reduce the impact of nonextreme observations. We also develop a local likelihood approach to capture the temporal dynamics of extremal dependence among pairs of leading cryptocurrencies. We here apply our model to historical closing prices of five leading cryotocurrencies which share large cryptocurrency market capitalizations. The results show that our proposed copula model outperforms alternative copula models and that the lower-tail dependence level between most pairs of leading cryptocurrencies and, in particular, Bitcoin and Ethereum has become stronger over time, smoothly transitioning from an asymptotic independence regime to an asymptotic dependence regime in recent years, whilst the upper tail has been relatively more stable overall at a weaker dependence level.

REFERENCES

ALCOCK, J. and SATCHELL, S. (2018). *Asymmetric Dependence in Finance*: *Diversification*, *Correlation and Portfolio Management in Market Downturns*. Wiley, New York.

ARELLANO-VALLE, R. B. and GENTON, M. G. (2010). Multivariate extended skew-*t* distributions and related families. *Metron* **68** 201–234. MR3041150 https://doi.org/10.1007/BF03263536

AULBACH, S., BAYER, V. and FALK, M. (2012). A multivariate piecing-together approach with an application to operational loss data. *Bernoulli* **18** 455–475. MR2922457 https://doi.org/10.3150/10-BEJ343

AZZALINI, A. and DALLA VALLE, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83** 715–726. MR1440039 https://doi.org/10.1093/biomet/83.4.715

BORRI, N. (2019). Conditional tail-risk in cryptocurrency markets. *J. Empir. Finance* **50** 1–19.

BROCKWELL, P. J. and DAVIS, R. A. (2002). *Introductin to Time Series and Forecasting*, 2nd ed. Springer, New York.

CASTRO-CAMILO, D., DE CARVALHO, M. and WADSWORTH, J. (2018). Time-varying extreme value dependence with application to leading European stock markets. *Ann. Appl. Stat.* **12** 283–309. MR3773394 https://doi.org/10.1214/17-AOAS1089

CASTRUCCIO, S., HUSER, R. and GENTON, M. G. (2016). High-order composite likelihood inference for max-stable distributions and processes. *J. Comput. Graph. Statist.* **25** 1212–1229. MR3572037 https://doi.org/10.1080/10618600.2015.1086656

Key words and phrases. Copula model, cryptocurrency data, extreme-value theory, joint-tail modeling, tail dependence.

COLES, S. G. and TAWN, J. A. (1991). Modelling extreme multivariate events. *J. Roy. Statist. Soc. Ser. B* **53** 377–392. MR1108334

DAVISON, A. C. and HUSER, R. (2015). Statistics of extremes. *Annu. Rev. Stat. Appl.* **2** 203–235.

DE CARVALHO, M., LEONELLI, M. and ROSSI, A. (2020). Tracking change-points in multivariate extremes. ArXiv preprint. Available at arXiv:2011.05067.

DE HAAN, L. and ZHOU, C. (2021). Trends in extreme value indices. *J. Amer. Statist. Assoc.* **116** 1265–1279. MR4309271 https://doi.org/10.1080/01621459.2019.1705307

DEMARTA, S. and MCNEIL, A. J. (2005). The t copula and related copulas. *Int. Stat. Rev.* **73** 111–129.

EINMAHL, J. H. J., DE HAAN, L. and ZHOU, C. (2016). Statistics of heteroscedastic extremes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 31–51. MR3453645 https://doi.org/10.1111/rssb.12099

EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (1997). *Modelling Extremal Events*: *For Insurance and Finance*. *Applications of Mathematics* (*New York*) **33**. Springer, Berlin. MR1458613 https://doi.org/10.1007/978-3-642-33483-2

ENGELKE, S., OPITZ, T. and WADSWORTH, J. (2019). Extremal dependence of random scale constructions. *Extremes* **22** 623–666. MR4031852 https://doi.org/10.1007/s10687-019-00353-3

FENG, W., WANG, Y. and ZHANG, Z. (2018). Can cryptocurrencies be a safe haven: A tail risk perspective analysis. *Appl. Econ.* **50** 4745–4762.

GKILLAS, K., BEKIROS, S. and SIRIOPOULOS, C. (2018). Extreme correlation in cryptocurrency markets. Available at SSRN 3180934.

GONG, Y. and HUSER, R. (2022a). Supplement to "Asymmetric tail dependence modeling, with application to cryptocurrency market data." https://doi.org/10.1214/21-AOAS1568SUPPA

GONG, Y. and HUSER, R. (2022b). R code for "Asymmetric tail dependence modeling, with application to cryptocurrency market data." https://doi.org/10.1214/21-AOAS1568SUPPB

HASHORVA, E. (2010). On the residual dependence index of elliptical distributions. *Statist. Probab. Lett.* **80** 1070–1078. MR2651047 https://doi.org/10.1016/j.spl.2010.03.001

HAZRA, A., REICH, B. J. and STAICU, A.-M. (2020). A multivariate spatial skew-t process for joint modeling of extreme precipitation indexes. *Environmetrics* **31** e2602.

HUSER, R., OPITZ, T. and THIBAUD, E. (2017). Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures. *Spat. Stat.* **21** 166–186. MR3692183 https://doi.org/10.1016/j.spasta.2017.06.004

HUSER, R. and WADSWORTH, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *J. Amer. Statist. Assoc.* **114** 434–444. MR3941266 https://doi.org/10.1080/01621459.2017.1411813

HUYNH, T. L. D., NGUYEN, S. P. and DUONG, D. (2018). Contagion risk measured by return among cryptocurrencies. In *International Econometric Conference of Vietnam* 987–998. Springer, Berlin.

KIRILIOUK, A., ROOTZÉN, H., SEGERS, J. and WADSWORTH, J. L. (2019). Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics* **61** 123–135. MR3933664 https://doi.org/10.1080/00401706.2018.1462738

KRUPSKII, P. (2017). Copula-based measures of reflection and permutation asymmetry and statistical tests. *Statist. Papers* **58** 1165–1187. MR3720957 https://doi.org/10.1007/s00362-016-0743-1

KRUPSKII, P., HUSER, R. and GENTON, M. G. (2018). Factor copula models for replicated spatial data. *J. Amer. Statist. Assoc.* **113** 467–479. MR3803479 https://doi.org/10.1080/01621459.2016.1261712

LEDFORD, A. W. and TAWN, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika* **83** 169–187. MR1399163 https://doi.org/10.1093/biomet/83.1.169

LEONELLI, M. and GAMERMAN, D. (2020). Semiparametric bivariate modelling with flexible extremal dependence. *Stat. Comput.* **30** 221–236. MR4064619 https://doi.org/10.1007/s11222-019-09878-w

MHALLA, L., DE CARVALHO, M. and CHAVEZ-DEMOULIN, V. (2019). Regression-type models for extremal dependence. *Scand. J. Stat.* **46** 1141–1167. MR4033807 https://doi.org/10.1111/sjos.12388

NAKAMOTO, S. (2008). Bitcoin: A peer-to-peer electronic cash system. Available at https://bitcoin.org/bitcoin.pdf.

NGUYEN, L. H., CHEVAPATRAKUL, T. and YAO, K. (2020). Investigating tail-risk dependence in the cryptocurrency markets: A LASSO quantile regression approach. *J. Empir. Finance* **58** 333–355.

PADOAN, S. A., RIBATET, M. and SISSON, S. A. (2010). Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc.* **105** 263–277. MR2757202 https://doi.org/10.1198/jasa.2009.tm08577

PATTON, A. J. (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *J. Financ. Econom.* **2** 130–168.

PATTON, A. J. (2006). Modelling asymmetric exchange rate dependence. *Internat. Econom. Rev.* **47** 527–556. MR2216591 https://doi.org/10.1111/j.1468-2354.2006.00387.x

POON, S.-H., ROCKINGER, M. and TAWN, J. (2003). Modelling extreme-value dependence in international stock markets. *Statist. Sinica* **13** 929–953. MR2026056

POON, S.-H., ROCKINGER, M. and TAWN, J. A. (2004). Extreme value dependence in financial markets: Diagnostics, models, financial implications. *Rev. Financ. Stud.* **17** 581–610.

ROOTZÉN, H., SEGERS, J. and WADSWORTH, J. L. (2018). Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *J. Multivariate Anal.* **165** 117–131. MR3768756 https://doi.org/10.1016/j.jmva.2017.12.003

ROOTZÉN, H. and TAJVIDI, N. (2006). Multivariate generalized Pareto distributions. *Bernoulli* **12** 917–930. MR2265668 https://doi.org/10.3150/bj/1161614952

SCARROTT, C. and MACDONALD, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT* **10** 33–60. MR2912370

SIBUYA, M. (1960). Bivariate extreme statistics. I. *Ann. Inst. Statist. Math. Tokyo* **11** 195–210. MR0115241 https://doi.org/10.1007/bf01682329

TAWN, J. A. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika* **75** 397–415. MR0967580 https://doi.org/10.1093/biomet/75.3.397

TAWN, J. A. (1990). Modelling multivariate extreme value distributions. *Biometrika* **77** 245–253.

VETTORI, S., HUSER, R. and GENTON, M. G. (2018). A comparison of dependence function estimators in multivariate extremes. *Stat. Comput.* **28** 525–538. MR3761339 https://doi.org/10.1007/s11222-017-9745-7

VRAC, M., NAVEAU, P. and DROBINSKI, P. (2007). Modeling pairwise dependencies in precipitation intensities. *Nonlinear Process. Geophys.* **14** 789–797.

WADSWORTH, J. L., TAWN, J. A., DAVISON, A. C. and ELTON, D. M. (2017). Modelling across extremal dependence classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 149–175. MR3597968 https://doi.org/10.1111/rssb.12157

ZHANG, Z., HUSER, R., OPITZ, T. and WADSWORTH, J. L. (2021). Modeling spatial extremes using normal mean-variance mixtures. *Extremes*. To appear.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1848–1867
https://doi.org/10.1214/21-AOAS1569

ANALYSIS OF PRESENCE-ONLY DATA VIA EXACT BAYES, WITH MODEL AND EFFECTS IDENTIFICATION

BY GUIDO A. MOREIRA[1,a] AND DANI GAMERMAN[2,b]

[1]*Centro de Biologia Molecular e Ambiental, Universidade do Minho,* [a]*guidoalber@gmail.com*
[2]*Instituto de Matemática, Universidade Federal do Rio de Janeiro,* [b]*dani@im.ufrj.br*

This paper provides an exact modeling approach for the analysis of presence-only ecological data. Our proposal is also based on frequently used inhomogeneous Poisson processes but does not rely on model approximations, unlike other approaches. Exactness is achieved via a data augmentation scheme. One of the augmented processes can be interpreted as the unobserved occurrences of the relevant species, and its posterior distribution can be used to make predictions of the species over the region of study beyond the observer bias. The data augmentation also leads to a natural Gibbs sampler to make Bayesian inference through MCMC. The proposal shows better performance than the currently standard method based on Poisson process with intensity function depending log-linearly on the covariates. Additionally, an identification problem that arises in the traditional model does not seem to affect our proposal in the analyses of real ecological data.

REFERENCES

ADAMS, R. P., MURRAY, I. and MACKAY, D. J. C. (2009). Tractable Nonparametric Bayesian Inference in Poisson Processes with Gaussian Process Intensities. In *Proceedings of the* 26*th Annual International Conference on Machine Learning*. *ICML'*09 9–16. Association for Computing Machinery, New York, NY, USA.

BADDELEY, A., RUBAK, E. and TURNER, R. (2015). *Spatial Point Patterns*: *Methodology and Applications with R*. CRC Press/CRC Press, London.

BYRNE, S. (2016). A note on the use of empirical AUC for evaluating probabilistic forecasts. *Electron. J. Stat.* **10** 380–393. MR3466187 https://doi.org/10.1214/16-EJS1109

CRESSIE, N. A. C. (1993). *Spatial Point Patterns*. Wiley, New York.

DIGGLE, P. J. (2014). *Statistical Analysis of Spatial and Spatio-temporal Point Patterns*, 3rd ed. *Monographs on Statistics and Applied Probability* **128**. CRC Press, Boca Raton, FL. MR3113855

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

DORAZIO, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Glob. Ecol. Biogeogr.* **23** 1472–1484.

ELITH, J. and LEATHWICK, J. (2007). Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions* **13** 265–275.

ELITH, J., GRAHAM, C., VALAVI, R., ABEGG, M., BRUCE, C., FORD, A., GUISAN, A., HIJMANS, R., HUETTMANN, F. et al. (2020). Presence-only and Presence-absence Data for Comparing Species Distribution Modeling Methods. *Biodiversity Informatics* **15** 69–80.

FIELDING, A. H. and BELL, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* **24** 38–49.

FITHIAN, W. and HASTIE, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *Ann. Appl. Stat.* **7** 1917–1939. MR3161707 https://doi.org/10.1214/13-AOAS667

FITHIAN, W., ELITH, J., HASTIE, T. and KEITH, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods Ecol. Evol.* **6** 424–438. https://doi.org/10.1111/2041-210X.12242

FLETCHER JR., R. J., HEFLEY, T. J., ROBERTSON, E. P., ZUCKERBERG, B., MCCLEERY, R. A. and DORAZIO, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology* **100** e02710.

Key words and phrases. Point process, presence-only, spatial statistics, Bayesian analysis, species distribution model.

GAMERMAN, D. and LOPES, H. F. (2006). *Markov Chain Monte Carlo*: *Stochastic simulation for Bayesian inference*, 2nd ed. *Texts in Statistical Science Series*. CRC Press/CRC, Boca Raton, FL. MR2260716

GELFAND, A. E. and SCHLIEP, E. M. (2018). *Bayesian Inference and Computing for Spatial Point Patterns*. *NSF-CBMS Regional Conference Series in Probability and Statistics* **10**. IMS, Beachwood, OH. MR3890052

GELFAND, A. E. and SHIROTA, S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecol. Monogr.* **89** e01372.

GELMAN, A., SIMPSON, D. and BETANCOURT, M. (2017). The prior can generally only be understood in the context of the likelihood.

GONÇALVES, F. B. and GAMERMAN, D. (2018). Exact Bayesian inference in spatiotemporal Cox processes driven by multivariate Gaussian processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 157–175. MR3744716 https://doi.org/10.1111/rssb.12237

HASTIE, T. and FITHIAN, W. (2013). Inference from presence-only data; the ongoing controversy. *Ecography* **36** 864–867. https://doi.org/10.1111/j.1600-0587.2013.00321.x

HEFLEY, T. J., TYRE, A. J., BAASCH, D. M. and BLANKENSHIP, E. E. (2013). Nondetection sampling bias in marked presence-only data. *Ecol. Evol.* **3** 5225–5236.

HEFLEY, T. J., BAASCH, D. M., TYRE, A. J. and BLANKENSHIP, E. E. (2015). Use of opportunistic sightings and expert knowledge to predict and compare Whooping Crane stopover habitat. *Conserv. Biol.* **29** 1337–1346.

JOURNÉ, V., BARNAGAUD, J.-Y., BERNARD, C., CROCHET, P.-A. and MORIN, X. (2020). Correlative climatic niche models predict real and virtual species distributions equally well. *Ecology* **101** e02912. https://doi.org/10.1002/ecy.2912

LEWIS, P. A. W. and SHEDLER, G. S. (1979). Simulation of nonhomogeneous Poisson processes by thinning. *Nav. Res. Logist. Q.* **26** 403–413. MR0546120 https://doi.org/10.1002/nav.3800260304

LITTLE, R. J. A. and RUBIN, D. (2014). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York.

MAZZOCHINI, G. G., FONSECA, C. R., COSTA, G. C., SANTOS, R. M., OLIVEIRA-FILHO, A. T. and GANADE, G. (2019). Plant phylogenetic diversity stabilizes large-scale ecosystem productivity. *Glob. Ecol. Biogeogr.* **28** 1430–1439.

MOREIRA, G. A. (2021). bayesPO: Bayesian Inference for Presence-Only Data. R package version 0.3.1.

MOREIRA, G. A. and GAMERMAN, D. (2022a). Supplement to "Analysis of presence-only data via exact Bayes, with model and effects identification." https://doi.org/10.1214/21-AOAS1569SUPPA

MOREIRA, G. A. and GAMERMAN, D. (2022b). Supplement to "Analysis of presence-only data via exact Bayes, with model and effects identification." https://doi.org/10.1214/21-AOAS1569SUPPB

OLIVEIRA-FILHO, A. T. (2017). NeoTropTree, Arborea flora of the Neotropical Region: A Database involving biogeography, diversity and consevation. Universidade Federal de Minas Gerais. Available at http://www.neotroptree.info.

PEARCE, J. L. and BOYCE, M. S. (2006). Modelling distribution and abundance with presence-only data. *J. Appl. Ecol.* **43** 405–412.

PHILLIPS, S. J., ANDERSON, R. P. and SCHAPIRE, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **190** 231–259.

PHILLIPS, S. J., DUDÍK, M. and SCHAPIRE, R. E. (2004). A Maximum Entropy Approach to Species Distribution Modeling. In *Proceedings of the Twenty-first International Conference on Machine Learning*. *ICML'*04 83. ACM, New York, NY, USA.

PHILLIPS, S. J., DUDÍK, M., ELITH, J., GRAHAM, C. H., LEHMANN, A., LEATHWICK, J. and FERRIER, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecol. Appl.* **19** 181–197.

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

RENNER, I. W., LOUVRIER, J. and GIMENEZ, O. (2019). Combining multiple data sources in species distribution models while accounting for spatial dependence and overfitting with combined penalised likelihood maximisation. BioRxiv.

RENNER, I. W. and WARTON, D. I. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics* **69** 274–281. MR3058074 https://doi.org/10.1111/j.1541-0420.2012.01824.x

RENNER, I. W., ELITH, J., BADDELEY, A., FITHIAN, W., HASTIE, T., PHILLIPS, S. J., POPOVIC, G. and WARTON, D. I. (2015). Point process models for presence-only analysis. *Methods Ecol. Evol.* **6** 366–379.

ROYLE, J. A., CHANDLER, R. B., YACKULIC, C. and NICHOLS, J. D. (2012). Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods Ecol. Evol.* **3** 545–554.

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

SHIROTA, S., GELFAND, A. E. and BANERJEE, S. (2019). Spatial joint species distribution modeling using Dirichlet processes. *Statist. Sinica* **29** 1127–1154. MR3932512

WARTON, D. I. and SHEPHERD, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Ann. Appl. Stat.* **4** 1383–1402. MR2758333 https://doi.org/10.1214/10-AOAS331

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1868–1890
https://doi.org/10.1214/21-AOAS1570

ESTIMATION OF THE MARGINAL EFFECT OF ANTIDEPRESSANTS ON BODY MASS INDEX UNDER CONFOUNDING AND ENDOGENOUS COVARIATE-DRIVEN MONITORING TIMES

BY JANIE COULOMBE[a], ERICA E. M. MOODIE[b], ROBERT W. PLATT[c] AND CHRISTEL RENOUX[d]

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, [a]*janie.coulombe@mail.mcgill.ca,* [b]*erica.moodie@mcgill.ca,* [c]*robert.platt@mcgill.ca,* [d]*christel.renoux@mcgill.ca*

In studying the marginal effect of antidepressants on body mass index using electronic health records data, we face several challenges. Patients' characteristics can affect the exposure (confounding) as well as the timing of routine visits (measurement process), and those characteristics may be altered following a visit which can create dependencies between the monitoring and body mass index when viewed as a stochastic or random processes in time. This may result in a form of selection bias that distorts the estimation of the marginal effect of the antidepressant. Inverse intensity of visit weights have been proposed to adjust for these imbalances, however no approaches have addressed complex settings where the covariate and the monitoring processes affect each other in time so as to induce endogeneity, a situation likely to occur in electronic health records. We review how selection bias due to outcome-dependent follow-up times may arise and propose a new cumulated weight that models a complete monitoring path so as to address the above-mentioned challenges and produce a reliable estimate of the impact of antidepressants on body mass index. More specifically, we do so using data from the Clinical Practice Research Datalink in the United Kingdom, comparing the marginal effect of two commonly used antidepressants, citalopram and fluoxetine, on body mass index. The results are compared to those obtained with simpler methods that do not account for the extent of the dependence due to an endogenous covariate process.

REFERENCES

ALAM, S., MOODIE, E. E. M. and STEPHENS, D. A. (2019). Should a propensity score model be super? The utility of ensemble procedures for causal adjustment. *Stat. Med.* **38** 1690–1702. MR3934814 https://doi.org/10.1002/sim.8075

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120. MR0673646

AUSTIN, P. C. and STUART, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat. Med.* **34** 3661–3679. MR3422140 https://doi.org/10.1002/sim.6607

BŮŽKOVÁ, P. and LUMLEY, T. (2009). Semiparametric modeling of repeated measurements under outcome-dependent follow-up. *Stat. Med.* **28** 987–1003. MR2518361 https://doi.org/10.1002/sim.3496

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

COOK, R. J. and LAWLESS, J. F. (2018). *Multistate Models for the Analysis of Life History Data. Monographs on Statistics and Applied Probability* **158**. CRC Press, Boca Raton, FL. MR3838371 https://doi.org/10.1201/9781315119731

COULOMBE, J., MOODIE, E. E. M. and PLATT, R. W. (2021). Weighted regression analysis to correct for informative monitoring times and confounders in longitudinal studies. *Biometrics* **77** 162–174. MR4229729 https://doi.org/10.1111/biom.13285

Key words and phrases. Body mass index, causal inference, citalopram, electronic health records, fluoxetine, outcome-dependent follow-up times.

COULOMBE, J., MOODIE, E. E. M., SHORTREED, S. and RENOUX, C. (2021). Can the risk of severe depression-related outcomes be reduced by tailoring the antidepressant therapy to patient characteristics? *Am. J. Epidemiol.* **190** 1210–1219.

COULOMBE, J., MOODIE, E. E., PLATT, R. W. and RENOUX, C. (2022). Supplement to "Estimation of the marginal effect of antidepressants on body mass index under confounding and endogenous covariate-driven monitoring times." https://doi.org/10.1214/21-AOAS1570SUPP

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

DEAS, I., ROBSON, B., WONG, C. and BRADFORD, M. (2003). Measuring neighbourhood deprivation: A critique of the index of multiple deprivation. *Environ. Plann., C, Gov. Policy* **21** 883–903.

DE LAS CUEVAS, C., PEÑATE, W. and SANZ, E. J. (2014). Risk factors for non-adherence to antidepressant treatment in patients with mood disorders. *Eur. J. Clin. Pharmacol.* **70** 89–98. https://doi.org/10.1007/s00228-013-1582-9

DIDELEZ, V. (2015). Causal reasoning for events in continuous time: A decision-theoretic approach. In *ACI@UAI* 40–45.

GAGNON, J., LUSSIER, M.-T., MACGIBBON, B., DASKALOPOULOU, S. S. and BARTLETT, G. (2018). The impact of antidepressant therapy on glycemic control in Canadian primary care patients with diabetes mellitus. *Front. Nutr.* **5** 47. https://doi.org/10.3389/fnut.2018.00047

GILL, R. D. and JOHANSEN, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Ann. Statist.* **18** 1501–1555. MR1074422 https://doi.org/10.1214/aos/1176347865

HERNÁN, M. A. and ROBINS, J. M. (2006). Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health* **60** 578–586.

HERRETT, E., GALLAGHER, A. M., BHASKARAN, K., FORBES, H., MATHUR, R., VAN STAA, T. and SMEETH, L. (2015). Data resource profile: Clinical practice research datalink (CPRD). *Int. J. Epidemiol.* **44** 827–836.

HU, L. and HOGAN, J. W. (2019). Causal comparative effectiveness analysis of dynamic continuous-time treatment initiation rules with sparsely measured outcomes and death. *Biometrics* **75** 695–707. MR3999191

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481. MR0093867

KYLE, R. P., MOODIE, E. E. M., KLEIN, M. B. and ABRAHAMOWICZ, M. (2019). Evaluating flexible modeling of continuous covariates in inverse-weighted estimators. *Am. J. Epidemiol.* **188** 1181–1191.

LIANG, Y., LU, W. and YING, Z. (2009). Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics* **65** 377–384. MR2751461 https://doi.org/10.1111/j.1541-0420.2008.01104.x

LIN, H., SCHARFSTEIN, D. O. and ROSENHECK, R. A. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 791–813. MR2088782 https://doi.org/10.1111/j.1467-9868.2004.b5543.x

LIN, D. Y., WEI, L. J., YANG, I. and YING, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 711–730. MR1796287 https://doi.org/10.1111/1467-9868.00259

LINDSEY, J. K. (2004). *Statistical Analysis of Stochastic Processes in Time*. *Cambridge Series in Statistical and Probabilistic Mathematics* **14**. Cambridge Univ. Press, Cambridge. MR2079123 https://doi.org/10.1017/CBO9780511617164

LIPSITZ, S. R., FITZMAURICE, G. M., IBRAHIM, J. G., GELBER, R. and LIPSHULTZ, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics* **58** 621–630. MR1933535 https://doi.org/10.1111/j.0006-341X.2002.00621.x

MOODIE, E. E. M. and STEPHENS, D. A. (2020). Comment: Clarifying endogeneous data structures and consequent modelling choices using causal graphs [MR4148211]. *Statist. Sci.* **35** 391–393. MR4148212 https://doi.org/10.1214/20-STS777

NEYMAN, J. S. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. MR1092986

PEARL, J. (2009). Causal inference in statistics: An overview. *Stat. Surv.* **3** 96–146. MR2545291 https://doi.org/10.1214/09-SS057

PETERSEN, M. L., PORTER, K. E., GRUBER, S., WANG, Y. and VAN DER LAAN, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* **21** 31–54. MR2867537 https://doi.org/10.1177/0962280210386207

PULLENAYEGUM, E. M. (2016). Multiple outputation for the analysis of longitudinal data subject to irregular observation. *Stat. Med.* **35** 1800–1818. MR3513486 https://doi.org/10.1002/sim.6829

PULLENAYEGUM, E. M. and LIM, L. S. H. (2016). Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Stat. Methods Med. Res.* **25** 2992–3014. MR3572895 https://doi.org/10.1177/0962280214536537

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

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

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

SERRETTI, A., MANDELLI, L., LAURA, M. et al. (2010). Antidepressants and body weight: A comprehensive review and meta-analysis. *J. Clin. Psychiatry* **71** 1259–1272.

STREETER, A. J., LIN, N. X., CRATHORNE, L., HAASOVA, M., HYDE, C., MELZER, D. and HENLEY, W. E. (2017). Adjusting for unmeasured confounding in nonrandomized longitudinal studies: A methodological review. *J. Clin. Epidemiol.* **87** 23–34.

SUSSMAN, N. and GINSBERG, D. (1998). Rethinking side effects of the selective serotonin reuptake inhibitors: Sexual dysfunction and weight gain. *Psychiatr. Ann.* **28** 89–97.

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

XIAO, Y., MOODIE, E. E. M. and ABRAHAMOWICZ, M. (2013). Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiol. Methods* **2** 1–20.

YANG, S., PIEPER, K. and COOLS, F. (2020). Semiparametric estimation of structural failure time models in continuous-time processes. *Biometrika* **107** 123–136. MR4064144 https://doi.org/10.1093/biomet/asz057

ZHU, Y., LAWLESS, J. F. and COTTON, C. A. (2017). Estimation of parametric failure time distributions based on interval-censored data with irregular dependent follow-up. *Stat. Med.* **36** 1548–1567. MR3631979 https://doi.org/10.1002/sim.7234

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1891–1918
https://doi.org/10.1214/21-AOAS1575

LARGE-SCALE MULTIVARIATE SPARSE REGRESSION WITH APPLICATIONS TO UK BIOBANK

BY JUNYANG QIAN[1,a], YOSUKE TANIGAWA[2,d], RUILIN LI[3,f], ROBERT TIBSHIRANI[1,b], MANUEL A. RIVAS[2,e] AND TREVOR HASTIE[1,c]

[1]*Department of Statistics, Stanford University,* [a]*junyangq@stanford.edu,* [b]*tibs@stanford.edu,* [c]*hastie@stanford.edu*
[2]*Department of Biomedical Data Science, Stanford University,* [d]*ytanigaw@stanford.edu,* [e]*mrivas@stanford.edu*
[3]*Institute for Computational and Mathematical Engineering, Stanford University,* [f]*ruilinli@stanford.edu*

In high-dimensional regression problems, often a relatively small subset of the features are relevant for predicting the outcome, and methods that impose sparsity on the solution are popular. When multiple correlated outcomes are available (multitask), reduced rank regression is an effective way to borrow strength and capture latent structures that underlie the data. Our proposal is motivated by the UK Biobank population-based cohort study, where we are faced with large-scale, ultrahigh-dimensional features, and have access to a large number of outcomes (phenotypes)—lifestyle measures, biomarkers, and disease outcomes. We are hence led to fit sparse reduced-rank regression models, using computational strategies that allow us to scale to problems of this size. We use a scheme that alternates between solving the sparse regression problem and solving the reduced rank decomposition. For the sparse regression component we propose a scalable iterative algorithm based on adaptive screening that leverages the sparsity assumption and enables us to focus on solving much smaller subproblems. The full solution is reconstructed and tested via an optimality condition to make sure it is a valid solution for the original problem. We further extend the method to cope with practical issues, such as the inclusion of confounding variables and imputation of missing values among the phenotypes. Experiments on both synthetic data and the UK Biobank data demonstrate the effectiveness of the method and the algorithm. We present `multiSnpnet` package, available at http://github.com/junyangq/multiSnpnet that works on top of `PLINK2` files, which we anticipate to be a valuable tool for generating polygenic risk scores from human genetic studies.

REFERENCES

ABADI, M., BARHAM, P., CHEN, J., CHEN, Z., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., IRVING, G. et al. (2016). **TensorFlow**: A system for large-scale machine learning. In *Proceedings of the* 12*th USENIX Conference on Operating Systems Design and Implementation*. *OSDI'*16 265–283. USENIX Association, Berkeley, CA, USA.

AGUIRRE, M., TANIGAWA, Y., VENKATARAMAN, G. R., TIBSHIRANI, R., HASTIE, T. and RIVAS, M. A. (2021). Polygenic risk modeling with latent trait-related genetic components. *Eur. J. Hum. Genet.*.

ANDERSON, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Stat.* **22** 327–351. MR0042664 https://doi.org/10.1214/aoms/1177729580

BACH, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9** 1179–1225. MR2417268

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

BOTTOU, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'*2010 177–186. Physica-Verlag/Springer, Heidelberg. MR3362066

BOVET, D. P. and CESATI, M. (2005). *Understanding the Linux Kernel*: *From I/O Ports to Process Management*. "O'Reilly Media, Inc."

Key words and phrases. Large-scale algorithm, ultrahigh-dimensional problem, sparse regression, reduced-rank regression, polygenic risk score, UK Biobank.

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

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

BYCROFT, C., FREEMAN, C., PETKOVA, D., BAND, G., ELLIOTT, L. T., SHARP, K., MOTYER, A., VUKCEVIC, D., DELANEAU, O. et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562** 203–209.

CHANG, C. C., CHOW, C. C., TELLIER, L. C., VATTIKUTI, S., PURCELL, S. M. and LEE, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4**.

CHEN, K. (2019). rrpack: Reduced-Rank Regression. R package version 0.1-11.

CHEN, L. and HUANG, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Amer. Statist. Assoc.* **107** 1533–1545. MR3036414 https://doi.org/10.1080/01621459.2012.734178

CHUN, H. and KELEŞ, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 3–25. MR2751241 https://doi.org/10.1111/j.1467-9868.2009.00723.x

COMON, P. (1994). Independent component analysis, a new concept? *Signal Process.* **36** 287–314.

DEAN, J. and GHEMAWAT, S. (2008). MapReduce: Simplified data processing on large clusters. *Commun. ACM* **51** 107–113.

DEBOEVER, C., TANIGAWA, Y., LINDHOLM, M. E., MCINNES, G., LAVERTU, A., INGELSSON, E., CHANG, C., ASHLEY, E. A., BUSTAMANTE, C. D. et al. (2018). Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* **9** 1612.

DUBOIS, B., DELMAS, J.-F. and OBOZINSKI, G. (2019). Fast algorithms for sparse reduced-rank regression. In *Proceedings of Machine Learning Research* (K. Chaudhuri and M. Sugiyama, eds.). *Proceedings of Machine Learning Research* **89** 2415–2424. PMLR.

DUCHI, J. C., AGARWAL, A. and WAINWRIGHT, M. J. (2012). Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Trans. Automat. Control* **57** 592–606. MR2932818 https://doi.org/10.1109/TAC.2011.2161027

EFRON, B. and HASTIE, T. (2016). *Computer Age Statistical Inference*: *Algorithms*, *Evidence*, *and Data Science*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **5**. Cambridge Univ. Press, New York. MR3523956 https://doi.org/10.1017/CBO9781316576533

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

GABRIEL, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58** 453–467. MR0312645 https://doi.org/10.1093/biomet/58.3.453

GOWER, J., LUBBE, S. and LE ROUX, N. (2011). *Understanding Biplots*. Wiley, Chichester. MR2829991 https://doi.org/10.1002/9780470973196

GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039 https://doi.org/10.3150/bj/1106314846

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika* **28** 321–377.

HYVÄRINEN, A. and OJA, E. (2000). Independent component analysis: Algorithms and applications. *Neural Netw.* **13** 411–430.

JUTTEN, C. and HERAULT, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Process.* **24** 1–10.

LELLO, L., AVERY, S. G., TELLIER, L., VAZQUEZ, A. I., DE LOS CAMPOS, G. and HSU, S. D. H. (2018). Accurate genomic prediction of human height. *Genetics* **210** 477–497. https://doi.org/10.1534/genetics.118.301267

LI, G., LIU, X. and CHEN, K. (2019). Integrative multi-view regression: Bridging group-sparse and low-rank models. *Biometrics* **75** 593–602. MR3999182 https://doi.org/10.1111/biom.13006

LI, R., CHANG, C., JUSTESEN, J. M., TANIGAWA, Y., QIANG, J., HASTIE, T., RIVAS, M. A. and TIBSHIRANI, R. (2020). Fast lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank. *Biostatistics*.

LIN, J., TABASSUM, R., RIPATTI, S. and PIRINEN, M. (2020). MetaPhat: Detecting and decomposing multivariate associations from univariate genome-wide association statistics. *Front. Genet.* **11** 431. https://doi.org/10.3389/fgene.2020.00431

LUO, C., LIANG, J., LI, G., WANG, F., ZHANG, C., DEY, D. K. and CHEN, K. (2018). Leveraging mixed and incomplete outcomes via reduced-rank modeling. *J. Multivariate Anal.* **167** 378–394. MR3830653 https://doi.org/10.1016/j.jmva.2018.04.011

MA, Z., MA, Z. and SUN, T. (2020). Adaptive estimation in two-way sparse reduced-rank regression. *Statist. Sinica* **30** 2179–2201. MR4260760 https://doi.org/10.5705/ss.20

MA, Z. and SUN, T. (2014). Adaptive sparse reduced-rank regression. ArXiv preprint. Available at arXiv:1403.1922.

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

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Statist.* **39** 1–47. MR2797839 https://doi.org/10.1214/09-AOS776

PENG, J., ZHU, J., BERGAMASCHI, A., HAN, W., NOH, D.-Y., POLLACK, J. R. and WANG, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.* **4** 53–77. MR2758084 https://doi.org/10.1214/09-AOAS271

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909. https://doi.org/10.1038/ng1847

QIAN, J., TANIGAWA, Y., DU, W., AGUIRRE, M., CHANG, C., TIBSHIRANI, R., RIVAS, M. A. and HASTIE, T. (2020). A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet.* **16** e1009141.

QIAN, J., TANIGAWA, Y., LI, R., TIBSHIRANI, R., RIVAS, M. A. and HASTIE, T. (2022). Supplement to "Large-scale multivariate sparse regression with applications to UK Biobank." https://doi.org/10.1214/21-AOAS1575SUPPA, https://doi.org/10.1214/21-AOAS1575SUPPB, https://doi.org/10.1214/21-AOAS1575SUPPC, https://doi.org/10.1214/21-AOAS1575SUPPD, https://doi.org/10.1214/21-AOAS1575SUPPE, https://doi.org/10.1214/21-AOAS1575SUPPF, https://doi.org/10.1214/21-AOAS1575SUPPG

REINSEL, G. C. and VELU, R. P. (1998). *Multivariate Reduced-Rank Regression*: *Theory and Applications*. *Lecture Notes in Statistics* **136**. Springer, New York. MR1719704 https://doi.org/10.1007/978-1-4757-2853-8

SHEN, L. and THOMPSON, P. M. (2020). Brain imaging genomics: Integrated analysis and machine learning. *Proc IEEE Inst Electr Electron Eng* **108** 125–162. https://doi.org/10.1109/JPROC.2019.2947272

SILVER, M., MONTANA, G. and INITIATIVE, A. D. N. (2012). Fast identification of biological pathways associated with a quantitative trait using group Lasso with overlaps. *Stat. Appl. Genet. Mol. Biol.* **11** Art. 7. MR2924204 https://doi.org/10.2202/1544-6115.1755

SILVER, M., JANOUSOVA, E., HUA, X., THOMPSON, P. M., MONTANA, G., INITIATIVE, A. D. N. et al. (2012). Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. *NeuroImage* **63** 1681–1694.

SIMILÄ, T. and TIKKA, J. (2007). Input selection and shrinkage in multiresponse linear regression. *Comput. Statist. Data Anal.* **52** 406–422. MR2409992 https://doi.org/10.1016/j.csda.2007.01.025

SINNOTT-ARMSTRONG, N., TANIGAWA, Y., AMAR, D., MARS, N., BENNER, C., AGUIRRE, M., VENKATARAMAN, G. R., WAINBERG, M., OLLILA, H. M. et al. (2021). Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53** 185–194.

TANIGAWA, Y., LI, J., JUSTESEN, J. M., HORN, H., AGUIRRE, M., DEBOEVER, C., CHANG, C., NARASIMHAN, B., LAGE, K. et al. (2019). Components of genetic associations across 2138 phenotypes in the UK Biobank highlight adipocyte biology. *Nat. Commun.* **10** 4064.

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

TIBSHIRANI, R., BIEN, J., FRIEDMAN, J., HASTIE, T., SIMON, N., TAYLOR, J. and TIBSHIRANI, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 245–266. MR2899862 https://doi.org/10.1111/j.1467-9868.2011.01004.x

TURLACH, B. A., VENABLES, W. N. and WRIGHT, S. J. (2005). Simultaneous variable selection. *Technometrics* **47** 349–363. MR2164706 https://doi.org/10.1198/004017005000000139

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

VOUNOU, M., NICHOLS, T. E. and MONTANA, G. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage* **53** 1147–1159. Imaging Genetics.

VOUNOU, M., JANOUSOVA, E., WOLZ, R., STEIN, J. L., THOMPSON, P. M., RUECKERT, D. and MONTANA, G. (2012). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *NeuroImage* **60** 700–716.

WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55** 2183–2202. MR2729873 https://doi.org/10.1109/TIT.2009.2016018

WOLD, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis* (*Proc. Internat. Sympos., Dayton, Ohio*, 1965) 391–420. Academic Press, New York. MR0220397

XIAO, L. (2010). Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.* **11** 2543–2596. MR2738777

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574 https://doi.org/10.1111/j.1467-9868.2005.00532.x

ZAHARIA, M., CHOWDHURY, M., FRANKLIN, M. J., SHENKER, S. and STOICA, I. (2010). **Spark**: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*. *HotCloud'*10 10–10. USENIX Association, Berkeley, CA, USA.

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

ZHU, X., ZHANG, W. and FAN, Y. (2018). A robust reduced rank graph regression method for neuroimaging genetic analysis. *Neuroinformatics* **16** 1–11.

ZHU, X., SUK, H.-I., HUANG, H. and SHEN, D. (2016). Structured sparse low-rank regression model for brain-wide and genome-wide associations. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI* 2016 (S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal and W. Wells, eds.) 344–352. Springer, Cham.

ZHU, X., SUK, H.-I., HUANG, H. and SHEN, D. (2017). Low-rank graph-regularized structured sparse regression for identifying genetic biomarkers. *IEEE Transactions on Big Data* **3** 405–414.

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

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1919–1936
https://doi.org/10.1214/21-AOAS1576

SPATIAL FUNCTIONAL DATA MODELING OF PLANT REFLECTANCES

BY PHILIP A. WHITE[1,a], HENRY FRYE[2,b], MICHAEL F. CHRISTENSEN[3,d], ALAN E. GELFAND[3,e] AND JOHN A. SILANDER, JR.[2,c]

[1]*Department of Statistics, Brigham Young University,* [a]*pwhite@stat.byu.edu*
[2]*Department of Ecology and Evolutionary Biology, University of Connecticut,* [b]*henry.frye@uconn.edu,* [c]*john.silander_jr@uconn.edu*
[3]*Department of Statistical Science, Duke University,* [d]*michael.f.christensen@duke.edu,* [e]*alan@duke.edu*

Plant reflectance spectra, the profile of light reflected by leaves across different wavelengths, supply the spectral signature for a species at a spatial location to enable estimation of functional and taxonomic diversity for plants. We consider leaf spectra as "responses" to be explained spatially. These reflectance spectra are also functions over wavelength that respond to the environment. Our motivating data are gathered for several plant families from the Greater Cape Floristic Region (GCFR) in South Africa and lead us to develop rich novel spatial models that can explain spectra for genera within families. Wavelength responses for an individual leaf are viewed as a function of wavelength, leading to functional data modeling. Local environmental features become covariates. We introduce a wavelength, covariate interaction, since the response to environmental regressors may vary with wavelength, as may variance. Formal spatial modeling enables prediction of reflectances for genera at unobserved locations with known environmental features. We incorporate spatial dependence, wavelength dependence, and space–wavelength interaction (in the spirit of space–time interaction). We implement out-of-sample validation for model selection, finding that the model features above are informative for the functional data analysis. We supply ecological interpretation of the results under the selected model.

REFERENCES

SHI, C. and WANG, L. (2014). Incorporating spatial information in spectral unmixing: A review. *Remote Sens. Environ.* **149** 70–87.

ASNER, G. P. and MARTIN, R. E. (2016). Spectranomics: Emerging science and conservation opportunities at the interface of biodiversity and remote sensing. *Global Ecology and Conservation* **8** 212–219.

ASNER, G. P., MARTIN, R. E., KNAPP, D. E., TUPAYACHI, R., ANDERSON, C. B., SINCA, F., VAUGHN, N. R. and LLACTAYO, W. (2017). Airborne laser-guided imaging spectroscopy to map forest trait diversity and guide conservation. *Science* **355** 385–389. https://doi.org/10.1126/science.aaj1987

BESSE, P. C., CARDOT, H. and STEPHENSON, D. B. (2000). Autoregressive forecasting of some functional climatic variations. *Scand. J. Stat.* **27** 673–687.

CAVENDER-BARES, J., MEIRELES, J. E., COUTURE, J. J., KAPROTH, M. A., KINGDON, C. C., SINGH, A., SERBIN, S. P., CENTER, A., ZUNIGA, E. et al. (2016). Associations of leaf spectra with genetic and phylogenetic variation in oaks: Prospects for remote detection of biodiversity. *Remote Sens.* **8** 221.

CAWSE-NICHOLSON, K. (2021). NASA's surface biology and geology designated observable: A perspective on surface imaging algorithms. *Remote Sens. Environ.* **257** 112349.

CLARK, M. L., ROBERTS, D. A. and CLARK, D. B. (2005). Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. *Remote Sens. Environ.* **96** 375–398.

CORNWELL, W. K., WESTOBY, M., FALSTER, D. S., FITZJOHN, R. G., O'MEARA, B. C., PENNELL, M. W., MCGLINN, D. J., EASTMAN, J. M., MOLES, A. T. et al. (2014). Functional distinctiveness of major plant lineages. *Journal of Ecology* **102** 345–356.

DOUGHTY, C. E. (2017). Can leaf spectroscopy predict leaf and forest traits along a Peruvian tropical forest elevation gradient?: Amazonian leaf spectroscopy and traits. *J. Geophys. Res., Biogeosci.* **122** 2952–2965.

Key words and phrases. Environmental regressors, functional data analysis, heterogeneity, hierarchical model, interaction, kernel weighting, reflectance, spatial confounding.

FENG, W., YAO, X., ZHU, Y., TIAN, Y. and CAO, W. (2008). Monitoring leaf nitrogen status with hyperspectral reflectance in wheat. *European Journal of Agronomy* **28** 394–404.

FÉRET, J. B. (2019). Estimating leaf mass per area and equivalent water thickness based on leaf optical properties: Potential and limitations of physical modeling and machine learning. *Remote Sens. Environ.* **231** 110959.

GAMON, J. A., WANG, R., GHOLIZADEH, H., ZUTTA, B., TOWNSEND, P. A. and CAVENDER-BARES, J. (2020). Consideration of scale in remote sensing of biodiversity. In *Remote Sensing of Plant Biodiversity* 425–447. Springer, Cham.

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

GUO, W. (2002). Functional mixed effects models. *Biometrics* **58** 121–128. MR1891050 https://doi.org/10.1111/j.0006-341X.2002.00121.x

HIGDON, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environ. Ecol. Stat.* **5** 173–190.

HIGDON, D. (2002). Space and space–time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* 37–56. Springer, London. MR2059819

HODGES, J. S. and REICH, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *Amer. Statist.* **64** 325–334. MR2758564 https://doi.org/10.1198/tast.2010.10052

JACQUEMOUD, S. and BARET, F. (1990). PROSPECT: A model of leaf optical properties spectra. *Remote Sens. Environ.* **34** 75–91.

JACQUEMOUD, S. and USTIN, S. (2019a). Modeling leaf optical properties: PROSPECT. In *Leaf Optical Properties* Cambridge Univ. Press, Cambridge.

JACQUEMOUD, S. and USTIN, S. (2019b). Variation due to leaf structural, chemical, and physiological traits. In *Leaf Optical Properties* 170–194. Cambridge Univ. Press, Cambridge.

JACQUEMOUD, S. and USTIN, S. (2019c). Leaf optical properties in different wavelength domains. In *Leaf Optical Properties* 124–169. Cambridge Univ. Press, Cambridge.

KHAN, K. and CALDER, C. A. (2020). Restricted spatial regression methods: Implications for inference. *J. Amer. Statist. Assoc.* 1–13.

KIMELDORF, G. S. and WAHBA, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat.* **41** 495–502. MR0254999 https://doi.org/10.1214/aoms/1177697089

KOKALY, R. F., ASNER, G. P., OLLINGER, S. V., MARTIN, M. E. and WESSMAN, C. A. (2009). Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies. *Remote Sens. Environ.* **113** S78–S91.

LAUKAITIS, A. (2008). Functional data analysis for cash flow and transactions intensity continuous-time prediction using Hilbert-valued autoregressive processes. *European J. Oper. Res.* **185** 1607–1614.

LENG, X. and MÜLLER, H.-G. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics* **22** 68–76. https://doi.org/10.1093/bioinformatics/bti742

LOCANTORE, N., MARRON, J. S., SIMPSON, D. G., TRIPOLI, N., ZHANG, J. T. and COHEN, K. L. (1999). Robust principal component analysis for functional data. *TEST* **8** 1–73. MR1707596 https://doi.org/10.1007/BF02595862

MORRIS, J. S. and CARROLL, R. J. (2006). Wavelet-based functional mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 179–199. MR2188981 https://doi.org/10.1111/j.1467-9868.2006.00539.x

ORDOÑEZ, C., MARTÍNEZ, J., MATÍAS, J. M., REYES, A. N. and RODRÍGUEZ-PÉREZ, J. R. (2010). Functional statistical techniques applied to vine leaf water content determination. *Math. Comput. Modelling* **52** 1116–1122. MR2718426 https://doi.org/10.1016/j.mcm.2010.03.008

QUINTANO, C., FERNÁNDEZ-MANSO, A., SHIMABUKURO, Y. E. and PEREIRA, G. (2012). Spectral unmixing. *Int. J. Remote Sens.* **33** 5307–5340.

RAMSAY, J. O. (1988). Monotone regression splines in action. *Statist. Sci.* 425–441.

RAMSAY, J. (2005). Functional data analysis. *Encyclopedia of Statistics in Behavioral Science*.

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

RAY, S. and MALLICK, B. (2006). Functional clustering by Bayesian wavelet methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 305–332. MR2188987 https://doi.org/10.1111/j.1467-9868.2006.00545.x

REICH, P. B., WRIGHT, I. J., CAVENDER-BARES, J., CRAINE, J., OLEKSYN, J., WESTOBY, M. and WALTERS, M. (2003). The evolution of plant functional variation: Traits, spectra, and strategies. *International Journal of Plant Sciences* **164** S143–S164.

REISS, P. T. and OGDEN, R. T. (2007). Functional principal component regression and functional partial least squares. *J. Amer. Statist. Assoc.* **102** 984–996. MR2411660 https://doi.org/10.1198/016214507000000527

SCHNEIDER, F. D., MORSDORF, F., SCHMID, B., PETCHEY, O. L., HUENI, A., SCHIMEL, D. S. and SCHAEPMAN, M. E. (2017). Mapping functional diversity from remotely sensed morphological and physiological forest traits. *Nat. Commun.* **8** 1–12.

SCHWEIGER, A. K., CAVENDER-BARES, J., TOWNSEND, P. A., HOBBIE, S. E., MADRITCH, M. D., WANG, R., TILMAN, D. and GAMON, J. A. (2018). Plant spectral diversity integrates functional and phylogenetic components of biodiversity and predicts ecosystem function. *Nat. Ecol. Evol.* **2** 976–982.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. MR1979380 https://doi.org/10.1111/1467-9868.00353

TANNER, M. A. (1996). *Tools for Statistical Inference*, 3rd ed. *Springer Series in Statistics*. Springer, New York. MR1396311 https://doi.org/10.1007/978-1-4612-4024-2

TAYLOR, H. C. (1996). *Cederberg Vegetation and Flora*, National Botanical Institute, Cape Town.

TIAN, P., TENG, I. C., MAY, L. D., KURZ, R., LU, K., SCADENG, M., HILLMAN, E. M., DE CRESPIGNY, A. J., D'ARCEUIL, H. E. et al. (2010). Cortical depth-specific microvascular dilation underlies laminar differences in blood oxygenation level-dependent functional MRI signal. *Proc. Natl. Acad. Sci. USA* **107** 15246–15251.

ULLAH, S. and FINCH, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC Med. Res. Methodol.* **13** 43. https://doi.org/10.1186/1471-2288-13-43

VAN DER MERWE, H., VAN ROOYEN, M. W., and VAN ROOYEN, N. (2008). Vegetation of the Hantam–Tanqua–Roggeveld subregion, South Africa part 2: Succulent karoo biome related vegetation. *Koedoe* **50** 160–183.

WACKERNAGEL, H. (1998). *Multivariate Geostatistics*. Springer, Berlin.

WHITE, P. A. and GELFAND, A. E. (2021). Multivariate functional data modeling with time-varying clustering. *TEST* **30** 586–602. MR4297269 https://doi.org/10.1007/s11749-020-00733-z

WHITE, P. A., KEELER, D. G. and RUPPER, S. (2021). Hierarchical integrated spatial process modeling of monotone West Antarctic snow density curves. *Ann. Appl. Stat.* **15** 556–571. MR4298972 https://doi.org/10.1214/21-aoas1443

WHITE, P. A., FRYE, H., CHRISTENSEN, M. F., GELFAND, A. E. and SILANDER, J. A. (2022). Supplement to "Spatial functional data modeling of plant reflectances." https://doi.org/10.1214/21-AOAS1576SUPP

YU, S., WANG, G., WANG, L., LIU, C. and YANG, L. (2020). Estimation and inference for generalized geoadditive models. *J. Amer. Statist. Assoc.* **115** 761–774. MR4107678 https://doi.org/10.1080/01621459.2019.1574584

ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261. MR2054303 https://doi.org/10.1198/016214504000000241

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1937–1961
https://doi.org/10.1214/21-AOAS1577

ICE MODEL CALIBRATION USING SEMICONTINUOUS SPATIAL DATA

BY WON CHANG[1,a], BLEDAR A. KONOMI[1,b], GEORGIOS KARAGIANNIS[2,c], YAWEN GUAN[3,d] AND MURALI HARAN[4,e]

[1]*Division of Statistics and Data Science, University of Cincinnati,* [a]*won.chang@uc.edu,* [b]*konomibr@ucmail.uc.edu*
[2]*Department of Mathematical Sciences, Durham University,* [c]*georgios.karagiannis@durham.ac.uk*
[3]*Department of Statistics, University of Nebraska,* [d]*yawen.guan@unl.edu*
[4]*Department of Statistics, Pennsylvania State University,* [e]*mharan@stat.psu.edu*

Rapid changes in Earth's cryosphere caused by human activity can lead to significant environmental impacts. Computer models provide a useful tool for understanding the behavior and projecting the future of Arctic and Antarctic ice sheets. However, these models are typically subject to large parametric uncertainties, due to poorly constrained model input parameters that govern the behavior of simulated ice sheets. Computer model calibration provides a formal statistical framework to infer parameters, using observational data, and to quantify the uncertainty in projections due to the uncertainty in these parameters. Calibration of ice sheet models is often challenging because the relevant model output and observational data take the form of semicontinuous spatial data with a point mass at zero and a right-skewed continuous distribution for positive values. Current calibration approaches cannot handle such data. Here, we introduce a hierarchical latent variable model that handles binary spatial patterns and positive continuous spatial patterns as separate components. To overcome challenges due to high dimensionality, we use likelihood-based generalized principal component analysis to impose low-dimensional structures on the latent variables for spatial dependence. We apply our methodology to calibrate a physical model for the Antarctic ice sheet and demonstrate that we can overcome the aforementioned modeling and computational challenges. As a result of our calibration, we obtain improved future ice-volume change projections.

REFERENCES

BAYARRI, M. J., BERGER, J. O., CAFEO, J., GARCIA-DONATO, G., LIU, F., PALOMO, J., PARTHASARATHY, R. J., PAULO, R., SACKS, J. et al. (2007). Computer model validation with functional output. *Ann. Statist.* **35** 1874–1906. MR2363956 https://doi.org/10.1214/009053607000000163

BERDAHL, M., LEGUY, G., LIPSCOMB, W. H. and URBAN, N. M. (2020). Statistical emulation of a perturbed basal melt ensemble of an ice sheet model to better quantify Antarctic sea level rise uncertainties. *Cryosphere* **15** 2683–2699.

BERGER, J. O., DE OLIVEIRA, V. and SANSÓ, B. (2001). Objective Bayesian analysis of spatially correlated data. *J. Amer. Statist. Assoc.* **96** 1361–1374. MR1946582 https://doi.org/10.1198/016214501753382282

BHAT, K. S., HARAN, M., OLSON, R. and KELLER, K. (2012). Inferring likelihoods and climate system characteristics from climate models and multiple tracers. *Environmetrics* **23** 345–362. MR2935569 https://doi.org/10.1002/env.2149

CAO, F., BA, S., BRENNEMAN, W. A. and JOSEPH, V. R. (2018). Model calibration with censored data. *Technometrics* **60** 255–262. MR3804253 https://doi.org/10.1080/00401706.2017.1345704

CHANG, W., HARAN, M., OLSON, R. and KELLER, K. (2014). Fast dimension-reduced climate model calibration and the effect of data aggregation. *Ann. Appl. Stat.* **8** 649–673. MR3262529 https://doi.org/10.1214/14-AOAS733

CHANG, W., HARAN, M., OLSON, R. and KELLER, K. (2015). A composite likelihood approach to computer model calibration with high-dimensional spatial data. *Statist. Sinica* **25** 243–259. MR3328813

Key words and phrases. Computer model emulation, computer model calibration, semicontinuous spatial data, West Antarctic ice sheet, dimension reduction.

CHANG, W., HARAN, M., APPLEGATE, P. and POLLARD, D. (2016a). Calibrating an ice sheet model using high-dimensional binary spatial data. *J. Amer. Statist. Assoc.* **111** 57–72. MR3494638 https://doi.org/10.1080/01621459.2015.1108199

CHANG, W., HARAN, M., APPLEGATE, P. and POLLARD, D. (2016b). Improving ice sheet model calibration using paleoclimate and modern data. *Ann. Appl. Stat.* **10** 2274–2302. MR3592057 https://doi.org/10.1214/16-AOAS979

CHANG, W., KONOMI, B. A., GEORGIOS, K., GUAN, Y. and HARAN, M. (2022). Supplement to "Ice model calibration using semicontinuous spatial data." https://doi.org/10.1214/21-AOAS1577SUPP

COOK, R. D. and NI, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Assoc.* **100** 410–428. MR2160547 https://doi.org/10.1198/016214504000001501

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

EDWARDS, T. L., BRANDON, M. A., DURAND, G., EDWARDS, N. R., GOLLEDGE, N. R., HOLDEN, P. B., NIAS, I. J., PAYNE, A. J., RITZ, C. and WERNECKE, A. (2019). Revisiting Antarctic ice loss due to marine ice-cliff instability. *Nature* **566** 58.

FRETWELL, P., PRITCHARD, H. D., VAUGHAN, D. G., BAMBER, J. L., BARRAND, N. E., BELL, R., BIANCHI, C., BINGHAM, R. G., BLANKENSHIP, D. D. et al. (2013). Bedmap2: Improved ice bed, surface and thickness datasets for Antarctica. *Cryosphere* **7** 375–393.

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

GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J., eds. (1995). *Markov Chain Monte Carlo in Practice. Interdisciplinary Statistics*. CRC Press, London. MR1397966 https://doi.org/10.1007/978-1-4899-4485-6

GLADSTONE, R. M., LEE, V., ROUGIER, J., PAYNE, A. J., HELLMER, H., LE BROCQ, A., SHEPHERD, A., EDWARDS, T. L., GREGORY, J. et al. (2012). Calibrated prediction of Pine Island Glacier retreat during the 21st and 22nd centuries with a coupled flowline model. *Earth Planet. Sci. Lett.* **333** 191–199.

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

GU, M. and BERGER, J. O. (2016). Parallel partial Gaussian process emulation for computer models with massive output. *Ann. Appl. Stat.* **10** 1317–1347. MR3553226 https://doi.org/10.1214/16-AOAS934

GU, M., PALOMO, J. and BERGER, J. O. (2019). RobustGaSP: Robust Gaussian stochastic process emulation in R. *R J.* **11** 112–136. https://doi.org/10.32614/RJ-2019-011

GU, M., WANG, X. and BERGER, J. O. (2018). Robust Gaussian stochastic process emulation. *Ann. Statist.* **46** 3038–3066. MR3851764 https://doi.org/10.1214/17-AOS1648

HARVILLE, D. A. (2008). *Matrix Algebra from a Statistician's Perspective*. Springer, Berlin.

HASTIE, T. J. (1992). Generalized additive models. In *Statistical Models in S* 249–307. Routledge, London.

HEATON, M. J., DATTA, A., FINLEY, A. O., FURRER, R., GUINNESS, J., GUHANIYOGI, R., GERBER, F., GRAMACY, R. B., HAMMERLING, D. et al. (2019). A case study competition among methods for analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* **24** 398–425.

HIGDON, D., GATTIKER, J., WILLIAMS, B. and RIGHTLEY, M. (2008). Computer model calibration using high-dimensional output. *J. Amer. Statist. Assoc.* **103** 570–583. MR2523994 https://doi.org/10.1198/016214507000000888

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

LE BROCQ, A. M., PAYNE, A. J. and VIELI, A. (2010). An improved Antarctic dataset for high resolution numerical ice sheet models (ALBMAP v1). *Earth Syst. Sci. Data* **2** 247–260.

LEE, S., HUANG, J. Z. and HU, J. (2010). Sparse logistic principal components analysis for binary data. *Ann. Appl. Stat.* **4** 1579–1601. MR2758342 https://doi.org/10.1214/10-AOAS327

LIU, Z., OTTO-BLIESNER, B., HE, F., BRADY, E., TOMAS, R., CLARK, P., CARLSON, A., LYNCH-STIEGLITZ, J., CURRY, W. et al. (2009). Transient simulation of last deglaciation with a new mechanism for Bølling–Allerød warming. *Science* **325** 310–314.

LOEPPKY, J. L., SACKS, J. and WELCH, W. J. (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics* **51** 366–376. MR2756473 https://doi.org/10.1198/TECH.2009.08040

POLLARD, D. and DECONTO, R. M. (2009). Modelling West Antarctic ice sheet growth and collapse through the past five million years. *Nature* **458** 329–332.

POLLARD, D. and DECONTO, R. M. (2012). Description of a hybrid ice sheet-shelf model, and application to Antarctica. *Geosci. Model Dev.* **5** 1273–1295.

POLLARD, D., DECONTO, R. M. and ALLEY, R. B. (2015). Potential Antarctic Ice Sheet retreat driven by hydrofracturing and ice cliff failure. *Earth Planet. Sci. Lett.* **412** 112–121.

POLLARD, D., CHANG, W., HARAN, M., APPLEGATE, P. and DECONTO, R. (2016). Large-ensemble modeling of last deglacial and future ice-sheet retreat in the Amundsen Sea Embayment, West Antarctica. *Geosci. Model Dev.* **9** 1697–1723.

SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. MR1041765

SALTER, J. M., WILLIAMSON, D. B., SCINOCCA, J. and KHARIN, V. (2019). Uncertainty quantification for computer models with spatial output using calibration-optimal bases. *J. Amer. Statist. Assoc.* **114** 1800–1814. MR4047301 https://doi.org/10.1080/01621459.2018.1514306

SANSÓ, B. and FOREST, C. (2009). Statistical calibration of climate system properties. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 485–503. MR2750089 https://doi.org/10.1111/j.1467-9876.2009.00669.x

STACKLIES, W., REDESTIG, H., SCHOLZ, M., WALTHER, D. and SELBIG, J. (2007). pcaMethods—A bioconductor package providing PCA methods for incomplete data. *Bioinformatics* **23** 1164–1167.

STEIN, M. L. (1992). Prediction and inference for truncated spatial data. *J. Comput. Graph. Statist.* **1** 91–110.

STONE, E. J., LUNT, D. J., RUTT, I. C. and HANNA, E. (2010). Investigating the sensitivity of numerical model simulations of the modern state of the Greenland ice-sheet and its future response to climate change. *Cryosphere* **4** 397–417.

SUNG, C.-L., HUNG, Y., RITTASE, W., ZHU, C. and WU, C. F. J. (2020). A generalized Gaussian process model for computer experiments with binary time series. *J. Amer. Statist. Assoc.* **115** 945–956. MR4107691 https://doi.org/10.1080/01621459.2019.1604361

TIPPING, M. E. and BISHOP, C. M. (1999). Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 611–622. MR1707864 https://doi.org/10.1111/1467-9868.00196

WOODBURY, M. A. (1950). *Inverting Modified Matrices*. Princeton Univ., Princeton, NJ. Statistical Research Group, Memo. Rep. no. 42. MR0038136

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1962–1985
https://doi.org/10.1214/21-AOAS1578

CAUSAL INFERENCE FOR TIME-VARYING TREATMENTS IN LATENT MARKOV MODELS: AN APPLICATION TO THE EFFECTS OF REMITTANCES ON POVERTY DYNAMICS

BY FEDERICO TULLIO[1,a] AND FRANCESCO BARTOLUCCI[2,b]

[1]*Directorate General for Economics, Statistics and Research, Bank of Italy,* [a]*federico.tullio@bancaditalia.it*
[2]*Department of Economics, University of Perugia,* [b]*francesco.bartolucci@unipg.it*

To assess the effectiveness of remittances on the poverty level of recipient households, we propose a causal inference approach that may be applied with longitudinal data and time-varying treatments. The method relies on the integration of a propensity score based technique, the inverse propensity weighting, with a general latent Markov (LM) framework. It is particularly useful when the outcome of interest is a characteristic that is not directly observable, and the analysis is focused on: (*i*) clustering units in a finite number of classes according to this latent characteristic and (*ii*) modelling the evolution of this characteristic across time depending on the received treatment. Parameter estimation is based on a two-step procedure. First, individual propensity score weights are computed accounting for predetermined covariates. Then, a weighted version of the standard LM model likelihood, based on such weights, is maximised by means of an expectation-maximisation algorithm or, alternatively, adopting a stepwise procedure. Finite-sample properties of the proposed estimators are studied by simulation. The application is focused on the effect of remittances on the poverty status of Ugandan households, based on a longitudinal survey spanning the period 2009–2014, and where manifest variables are indicators of deprivation. We find that remittances reduce the probability of falling into poverty, whereas they exert no impact on the probability of moving out of poverty.

REFERENCES

ADAMS JR., R. H. (2011). Evaluating the economic impact of international remittances on developing countries using household surveys: A literature review. *J. Dev. Stud.* **47** 809–828.

ADAMS JR., R. H. and CUECUECHA, A. (2010). Remittances, household expenditure and investment in Guatemala. *World Dev.* **38** 1626–1641.

ALKIRE, S. and FOSTER, J. (2011a). Counting and multidimensional poverty measurement. *J. Public Econ.* **95** 476–487.

ALKIRE, S. and FOSTER, J. (2011b). Understandings and misunderstandings of multidimensional poverty measurement. *J. Econ. Inequal.* **9** 289–314.

AMUEDO-DORANTES, C. and POZO, S. (2011). Remittances and income smoothing. *Am. Econ. Rev.* **101** 582–587.

ANTÓN, J. I. (2010). The impact of remittances on nutritional status of children in Ecuador. *Int. Migr. Rev.* **44** 269–299.

ASKAROV, Z. and DOUCOULIAGOS, H. (2020). A meta-analysis of the effects of remittances on household education expenditure. *World Dev.* **129** 104860.

ATKINSON, A. B. (2003). Multidimensional deprivation: Contrasting social welfare and counting approaches. *J. Econ. Inequal.* **1** 51–65.

BARTOLUCCI, F., FARCOMENI, A. and PENNONI, F. (2013). *Latent Markov Models for Longitudinal Data*. CRC Press, Boca Raton, FL. MR3184304

BARTOLUCCI, F., FARCOMENI, A. and PENNONI, F. (2014). Latent Markov models: A review of a general framework for the analysis of longitudinal data with covariates. *TEST* **23** 433–465. MR3252086 https://doi.org/10.1007/s11749-014-0381-7

Key words and phrases. Causal inference, expectation-maximisation algorithm, potential outcomes, weighted maximum likelihood.

BARTOLUCCI, F., MONTANARI, G. E. and PANDOLFI, S. (2015). Three-step estimation of latent Markov models with covariates. *Comput. Statist. Data Anal.* **83** 287–301. MR3281812 https://doi.org/10.1016/j.csda.2014.10.017

BARTOLUCCI, F., PANDOLFI, S. and PENNONI, F. (2017). LMest: An R package for latent Markov models for longitudinal categorical data. *J. Stat. Softw.* **81** 1–38.

BARTOLUCCI, F., PENNONI, F. and VITTADINI, G. (2016). Causal latent Markov model for the comparison of multiple treatments in observational longitudinal studies. *J. Educ. Behav. Stat.* **41** 146–179.

BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* **41** 164–171. MR0287613 https://doi.org/10.1214/aoms/1177697196

BOURGUIGNON, F. and CHAKRAVARTY, S. R. (2003). The measurement of multidimensional poverty. *J. Econ. Inequal.* **1** 25–49.

CHOWDHURY, M. and RADICIC, D. (2019). Remittances and asset accumulation in Bangladesh: A study using generalised propensity score. *J. Int. Dev.* **31** 475–494.

COX-EDWARDS, A. and RODRÍGUEZ-OREGGIA, E. (2009). Remittances and labor force participation in Mexico: An analysis using propensity score matching. *World Dev.* **37** 1004–1014.

DE HAAS, H. (2010). Migration and development: A theoretical perspective. *Int. Migr. Rev.* **44** 227–264.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DÉMURGER, S. (2015). Migration and families left behind. *IZA World Labor* **144**.

DEVEREUX, S. and SABATES-WHEELER, R. (2004). Transformative social protection. Technical Report 232, IDS.

DI MARI, R. and BAKK, Z. (2018). Mostly harmless direct effects: A comparison of different latent Markov modeling approaches. *Struct. Equ. Model.* **25** 467–483. MR3782743 https://doi.org/10.1080/10705511.2017.1387860

DI MARI, R., OBERSKI, D. L. and VERMUNT, J. K. (2016). Bias-adjusted three-step latent Markov modeling with covariates. *Struct. Equ. Model.* **23** 649–660. MR3548105 https://doi.org/10.1080/10705511.2016.1191015

DOTTO, F., FARCOMENI, A., PITTAU, M. G. and ZELLI, R. (2019). A dynamic inhomogeneous latent state model for measuring material deprivation. *J. Roy. Statist. Soc. Ser. A* **182** 495–516. MR3902669 https://doi.org/10.1111/rssa.12408

EFRON, B. and TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap*. **57**. CRC Press, New York. MR1270903 https://doi.org/10.1007/978-1-4899-4541-9

ELZE, M. C., GREGSON, J., BABER, U., WILLIAMSON, E., SARTORI, S., MEHRAN, R., NICHOLS, M., STONE, G. W. and POCOCK, S. J. (2017). Comparison of propensity score methods and covariate adjustment: Evaluation in 4 cardiovascular studies. *J. Am. Coll. Cardiol.* **69** 345–357. https://doi.org/10.1016/j.jacc.2016.10.060

FARCOMENI, A., RANALLI, M. and VIVIANI, S. (2021). Dimension reduction for longitudinal multivariate data by optimizing class separation of projected latent Markov models. *TEST* **30** 462–480. MR4265968 https://doi.org/10.1007/s11749-020-00727-x

GARCÍA-ESCUDERO, L. A., GORDALIZA, A., MATRÁN, C. and MAYO-ISCAR, A. (2015). Avoiding spurious local maximizers in mixture modeling. *Stat. Comput.* **25** 619–633. MR3334421 https://doi.org/10.1007/s11222-014-9455-3

GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231. MR0370936 https://doi.org/10.1093/biomet/61.2.215

HECKMAN, J. J., ICHIMURA, H. and TODD, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Rev. Econ. Stud.* **6** 605–654.

HERNÁN, M. A. and ROBINS, J. M. (2020). *Causal Inference*: *What If*. CRC Press/CRC Press, Boca Raton, FL.

ICHINO, A., MEALLI, F. and NANNICINI, T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *J. Appl. Econometrics* **23** 305–327. MR2420362 https://doi.org/10.1002/jae.998

IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Amer. Econ. Rev.* **93** 126–132.

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

KRISHNAKUMAR, J. (2008). Multidimensional measures of poverty and well-being based on latent variable models. In *Quantitative Approaches to Multidimensional Poverty Measurement* (N. Kakwani and J. Silver, eds.) 118–134, Chapter 7. Palgrave Macmillan, London, UK.

KUROSAKI, T. (2006). Consumption vulnerability to risk in rural Pakistan. *J. Dev. Stud.* **42** 70–89.

LANZA, S. T., COFFMAN, D. L. and XU, S. (2013). Causal inference in latent class analysis. *Struct. Equ. Model.* **20** 361–383. MR3259688 https://doi.org/10.1080/10705511.2013.797816
LANZA, S. T. and COLLINS, L. M. (2008). A new SAS procedure for latent transition analysis: Transitions in dating and sexual behavior. *Dev. Psychol.* **44** 446–456.
LAZARSFELD, P. F. and HENRY, N. W. (1969). *Latent Structure Analysis*. Houghton Mifflin, Boston, MA.
LEROUX, B. G. and PUTERMAN, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* **48** 545–558.
LUCAS, R. E. and STARK, O. (1985). Motivations to remit: Evidence from Botswana. *J. Polit. Econ.* **93** 901–918.
LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **23** 2937–2960.
MACHADO, C., PAULINO, C. D. and NUNES, F. (2009). Deprivation analysis based on Bayesian latent class models. *J. Appl. Stat.* **36** 871–891. MR2750730 https://doi.org/10.1080/02664760802520769
MARUOTTI, A. and PUNZO, A. (2021). Initialization of hidden Markov and semi-Markov models: A critical evaluation of several strategies. *Int. Stat. Rev.* **89** 447–480.
MBAYE, M. L. (2021). Remittances and rural credit markets: Evidence from Senegal. *Rev. Dev. Econ.* **25** 183–199.
MCCAFFREY, D. F., GRIFFIN, B. A., ALMIRALL, D., SLAUGHTER, M. E., RAMCHAND, R. and BURGETTE, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat. Med.* **32** 3388–3414. MR3074364 https://doi.org/10.1002/sim.5753
MCFADDEN, D. (1973). Conditional logit analysis of qualitative choice behaviour. In *Frontiers in Econometrics* (P. Zarembka, ed.) 105–142. Academic Press, New York.
MCKENZIE, D. and SASIN, M. J. (2007). Migration, remittances, poverty, and human capital: Conceptual and empirical challenges. Technical Report 4272, World Bank, Washington, D.C.
MCKENZIE, D. and YANG, D. (2012). Experimental approaches in migration studies. In *Handbook of Research Methods in Migration*, Chapter 12 (C. Vargas-Silva, ed.). Edward Elgar Publishing.
MOISIO, P. (2004). A latent class application to the multidimensional measurement of poverty. *Qual. Quant.* **38** 703–717.
MOLINA, J., SUED, M. and VALDORA, M. (2018). Models for the propensity score that contemplate the positivity assumption and their application to missing data and causality. *Stat. Med.* **37** 3503–3518. MR3862900 https://doi.org/10.1002/sim.7827
MUNSHI, K. (2003). Networks in the modern economy: Mexican migrants in the U.S. labor market. *Q. J. Econ.* **118** 549–599.
RAMASWAMY, V., DESARBO, W. S., REIBSTEIN, D. J. and ROBINSON, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Mark. Sci.* **12** 103–124.
RAPOPORT, H. and DOCQUIER, F. (2006). The economics of migrants' remittances. In *Handbook of the Economics of Giving, Altruism and Reciprocity*, 1st ed. (S. Kolm and J. M. Ythier, eds.) **1** 1135–1198, Chapter 17. Elsevier, Amsterdam.
RATHA, D., MOHAPATRA, S. and SCHEJA, E. (2011). Impact of migration on economic and social development: A review of evidence and emerging issues. Technical Report 5558, World Bank, Washington, D.C.
ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512. MR0877758 https://doi.org/10.1016/0270-0255(86)90088-6
ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.
ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90** 106–121. MR1325118
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
RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.
RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071 https://doi.org/10.1198/016214504000001880
SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014
SEN, A. (1976). Poverty: An ordinal approach to measurement. *Econometrica* **44** 219–231. MR0443901 https://doi.org/10.2307/1912718
SEN, A. K. (1980). Equality of what? In *The Tanner Lecture on Human Values* 197–220. Cambridge Univ. Press, Cambridge, UK.
SEN, A. K. (1981). *Poverty and Famines*: *An Essay on Entitlement and Deprivation*. The Clarendon Press, Oxford, UK.

STARK, O. and BLOOM, D. E. (1985). The new economics of labor migration. *Am. Econ. Rev.* **75** 173–178.
TOWNSEND, P. (1979). *Poverty in the United Kingdom*. Allen Lane and Penguin Books, London, UK.
TOWNSEND, P. (1987). Deprivation. *J. Soc. Policy* **16** 125–146.
TSIMPO NKENGNE, C. (2016). *The Uganda Poverty Assessment Report* 2016. World Bank Group, Washington, D.C.
TULLIO, F. and BARTOLUCCI, F. (2022). Supplement to "Causal inference for time-varying treatments in latent Markov models: An application to the effects of remittances on poverty dynamics." https://doi.org/10.1214/21-AOAS1578SUPP
UGANDA BUREAU OF STATISTICS AND BANK OF UGANDA (2008). *Uganda*: *Workers' Remittances Report*: *Inwards Remittances* 2006.
UNDP (2016). *Human Development Report*. United Nations Development Programme, New York, NY.
VERMUNT, J. K., LANGEHEINE, R. and BOCKENHOLT, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *J. Educ. Behav. Stat.* **24** 179–207.
WELCH, L. R. (2003). Hidden Markov models and the Baum–Welch algorithm. *IEEE Inf. Theory Soc. Newsl.* **53** 1–13.
WFP (2008). *Technical Guidance Sheet—Food Consumption Analysis*: *Calculation and Use of the Food Consumption Score in Food Security Analysis*. World Food Programme, Rome, Italy.
WHELAN, C. T. and MAITRE, B. (2006). Comparing poverty and deprivation dynamics: Issues of reliability and validity. *J. Econ. Inequal.* **4** 303–323.
YANG, D. (2008). International migration, remittances and household investment: Evidence from Philippine migrants' exchange rate shocks. *Econ. J.* **118** 591–630.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 1986–2009
https://doi.org/10.1214/21-AOAS1579

HETEROGENEOUS CAUSAL EFFECTS WITH IMPERFECT COMPLIANCE: A BAYESIAN MACHINE LEARNING APPROACH

BY FALCO J. BARGAGLI-STOFFI[1,a], KRISTOF DE WITTE[2,b] AND GIORGIO GNECCO[3,c]

[1]*Department of Biostatistics, Harvard T.H. Chan School of Public Health,* [a]*fbargaglistoffi@hsph.harvard.edu*
[2]*Faculty of Economics and Business, KU Leuven,* [b]*kristof.dewitte@kuleuven.be*
[3]*AXES Research Unit, IMT School for Advanced Studies,* [c]*giorgio.gnecco@imtlucca.it*

This paper introduces an innovative Bayesian machine learning algorithm to draw interpretable inference on heterogeneous causal effects in the presence of imperfect compliance (e.g., under an irregular assignment mechanism). We show, through Monte Carlo simulations, that the proposed Bayesian Causal Forest with Instrumental Variable (BCF-IV) methodology outperforms other machine learning techniques tailored for causal inference in discovering and estimating the heterogeneous causal effects while controlling for the familywise error rate (or, less stringently, for the false discovery rate) at leaves' level. BCF-IV sheds a light on the heterogeneity of causal effects in instrumental variable scenarios and, in turn, provides the policymakers with a relevant tool for targeted policies. Its empirical application evaluates the effects of additional funding on students' performances. The results indicate that BCF-IV could be used to enhance the effectiveness of school funding on students' performance.

REFERENCES

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

ANGRIST, J. D. and KRUEGER, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *J. Econ. Perspect.* **15** 69–85.

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

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

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

BALKE, A. and PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *J. Amer. Statist. Assoc.* **92** 1171–1176.

BARGAGLI-STOFFI, F. J. and GNECCO, G. (2018). Estimating heterogeneous causal effects in the presence of irregular assignment mechanisms. In *Proceedings of the* 5*th IEEE Conference in Data Science and Advanced Analytics* 1–10.

BARGAGLI-STOFFI, F. J., DE-WITTE, K. and GNECCO, G. (2022). Supplement to "Heterogeneous causal effects with imperfect compliance: A Bayesian machine learning approach." https://doi.org/10.1214/21-AOAS1579SUPPA, https://doi.org/10.1214/21-AOAS1579SUPPB

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

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

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

BREIMAN, L. (1984). *Classification and Regression Trees*. Routledge, New York.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. *Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

Key words and phrases. Causal inference, instrumental variable, heterogeneous effects, interpretable machine learning, school funding, students' performance.

CALONICO, S., CATTANEO, M. D. and TITIUNIK, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* **82** 2295–2326. MR3301169 https://doi.org/10.3982/ECTA11757

CALONICO, S., CATTANEO, M. D. and TITIUNIK, R. (2015). Rdrobust: An R package for robust nonparametric inference in regression-discontinuity designs. *R J.* **7** 38–51.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172 https://doi.org/10.1214/09-AOAS285

COLEMAN, J. S. (1966). Equality of educational opportunity. US Government Printing Office, Washington, DC, 1–32.

COOK, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *J. Econometrics* **142** 636–654. MR2416822 https://doi.org/10.1016/j.jeconom.2007.05.002

COOK, D. I., GEBSKI, V. J. and KEECH, A. C. (2004). Subgroup analysis in clinical trials. *Med. J. Aust.* **180** 289–291.

D'INVERNO, G., SMET, M. and DE WITTE, K. (2021). Impact evaluation in a multi-input multi-output setting: Evidence on the effect of additional resources for schools. *European J. Oper. Res.* **290** 1111–1124.

DOMINICI, F., BARGAGLI-STOFFI, F. J. and MEALLI, F. (2021). From controlled to undisciplined data: Estimating causal effects in the era of data science using a potential outcome framework. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8102afed

DORIE, V., HILL, J. and DORIE, M. V. (2020). Package 'bartCause'.

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

EBERTS, R. W. and STONE, J. A. (1988). Student achievement in public schools: Do principals make a difference? *Econ. Educ. Rev.* **7** 291–299.

GENTILUCCI, J. L. and MUTO, C. C. (2007). Principals' influence on academic achievement: The student perspective. *NASSP Bull.* **91** 219–236.

GOLDHABER, D. and HANSEN, M. (2010). Using performance on the job to inform teacher tenure decisions. *Am. Econ. Rev.* **100** 250–55.

HAHN, P. R., DORIE, V. and MURRAY, J. S. (2019). Atlantic causal inference conference (ACIC) data analysis challenge 2017. arXiv preprint arXiv:1905.09515.

HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal.* **15** 965–1056. Includes comments and discussions by 25 discussants and a rejoinder by the authors. MR4154846 https://doi.org/10.1214/19-BA1195

HAHN, J., TODD, P. and VAN DER KLAAUW, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69** 201–209.

HAHN, P. R., CARVALHO, C. M., PUELZ, D. and HE, J. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal.* **13** 163–182. MR3737947 https://doi.org/10.1214/16-BA1044

HANUSHEK, E. A. (2003). The failure of input-based schooling policies. *Econ. J.* **113** F64–F98.

HANUSHEK, E. A., MACHIN, S. J. and WOESSMANN, L. (2016). *Handbook of the Economics of Education*. Elsevier.

HANUSHEK, E. A. and WOESSMANN, L. (2017). School resources and student achievement: A review of cross-country economic research. In *Cognitive Abilities and Educational Outcomes* 149–171. Springer.

HARRIS, D. N. and SASS, T. R. (2011). Teacher training, teacher quality and student achievement. *J. Public Econ.* **95** 798–812.

HARTFORD, J., LEWIS, G., LEYTON-BROWN, K. and TADDY, M. (2017). Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning* 1414–1423. PMLR.

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

HOCHBERG, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75** 800–802. MR0995126 https://doi.org/10.1093/biomet/75.4.800

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6** 65–70. MR0538597

HOLMLUND, H. and SUND, K. (2008). Is the gender gap in school performance affected by the sex of the teacher? *Labour Econ.* **15** 37–53.

HOMMEL, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75** 383–386.

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

IMBENS, G. W. and LEMIEUX, T. (2008). Regression discontinuity designs: A guide to practice. *J. Econometrics* **142** 615–635. MR2416821 https://doi.org/10.1016/j.jeconom.2007.05.001

IMBENS, G. W. and RUBIN, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *Rev. Econ. Stud.* **64** 555–574. MR1485828 https://doi.org/10.2307/2971731

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

JACKSON, C. K. (2018). Does school spending matter? The new literature on an old question. Technical report, National Bureau of Economic Research.

JACKSON, C. K., JOHNSON, R. C. and PERSICO, C. (2015). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *Q. J. Econ.* **131** 157–218.

JOHNSON, M., CAO, J. and KANG, H. (2019). Detecting heterogeneous treatment effect with instrumental variables. arXiv preprint arXiv:1908.03652.

KAPELNER, A. and BLEICH, J. (2013). BartMachine: Machine learning with Bayesian additive regression trees. arXiv preprint arXiv:1312.2171.

KIM, B., KHANNA, R. and KOYEJO, O. O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. In *Advances in Neural Information Processing Systems* 2280–2288.

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

LEE, D. S. and LEMIEUX, T. (2010). Regression discontinuity designs in economics. *J. Econ. Lit.* **48** 281–355.

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

LI, F., MATTEI, A. and MEALLI, F. (2015). Evaluating the causal effect of university grants on student dropout: Evidence from a regression discontinuity design using principal stratification. *Ann. Appl. Stat.* **9** 1906–1931. MR3456358 https://doi.org/10.1214/15-AOAS881

LOGAN, B. R., SPARAPANI, R., MCCULLOCH, R. E. and LAUD, P. W. (2019). Decision making and uncertainty quantification for individualized treatments using Bayesian additive regression trees. *Stat. Methods Med. Res.* **28** 1079–1093. MR3934636 https://doi.org/10.1177/0962280217746191

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

MATTEI, A. and MEALLI, F. (2016). Regression discontinuity designs as local randomized experiments. *Observational Studies* **66** 156–173.

MEALLI, F. and RAMPICHINI, C. (2012). Evaluating the effects of university grants by using regression discontinuity designs. *J. Roy. Statist. Soc. Ser. A* **175** 775–798. MR2948374 https://doi.org/10.1111/j.1467-985X.2011.01022.x

MILLER, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* **267** 1–38. MR3874511 https://doi.org/10.1016/j.artint.2018.07.007

NOVAK, R., BAHRI, Y., ABOLAFIA, D. A., PENNINGTON, J. and SOHL-DICKSTEIN, J. (2018). Sensitivity and generalization in neural networks: An empirical study. arXiv preprint arXiv:1802.08760.

OECD (2017). Educational opportunity for all: Overcoming inequality throughout the life course.

PROSPERI, M., GUO, Y., SPERRIN, M., KOOPMAN, J. S., MIN, J. S., HE, X., RICH, S., WANG, M., BUCHAN, I. E. et al. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* **2** 369–375.

PSACHAROPOULOS, G. and PATRINOS, H. A. (2018). Returns to investment in education: A decennial review of the global literature. *Education Economics* **26** 445–458.

RICE, J. K. (2010). The impact of teacher experience: Examining the evidence and policy implications. National Center for Analysis of Longitudinal Data in Education Research.

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

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

RUBIN, D. B. (1986). Comment: Which ifs have causal answers. *J. Amer. Statist. Assoc.* **81** 961–962.

SARKAR, S. K. and CHANG, C.-K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Amer. Statist. Assoc.* **92** 1601–1608. MR1615269 https://doi.org/10.2307/2965431

THERNEAU, T., ATKINSON, B., RIPLEY, B. and RIPLEY, M. B. (2015). Package 'rpart' package version 4.1-15. Available at: https://cran.r-project.org/web/packages/rpart/rpart.pdf.

THISTLETHWAITE, D. L. and CAMPBELL, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *J. Educ. Psychol.* **51** 309.

TROCHIM, W. M. (1984). *Research Design for Program Evaluation*: *The Regression-Discontinuity Approach* **6**. SAGE Publications, Inc.

WANG, G., LI, J. and HOPP, W. J. (2018). An instrumental variable tree approach for detecting heterogeneous treatment effects in observational studies. Available at SSRN 3045327.

WENDLING, T., JUNG, K., CALLAHAN, A., SCHULER, A., SHAH, N. H. and GALLEGO, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Stat. Med.* **37** 3309–3324. MR3856345 https://doi.org/10.1002/sim.7820

WOODY, S., CARVALHO, C. M., HAHN, P. R. and MURRAY, J. S. (2020). Estimating heterogeneous effects of continuous exposures using Bayesian tree ensembles: Revisiting the impact of abortion rates on crime. arXiv preprint arXiv:2007.09845.

YEKUTIELI, D. (2008). Hierarchical false discovery rate-controlling methodology. *J. Amer. Statist. Assoc.* **103** 309–316. MR2420235 https://doi.org/10.1198/016214507000001373

ZHANG, H. and SINGER, B. H. (2010). *Recursive Partitioning and Applications*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2674991 https://doi.org/10.1007/978-1-4419-6824-1

ZHANG, Y. and WALLACE, B. C. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (*Volume* 1: *Long Papers*) 253–263.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 2010–2029
https://doi.org/10.1214/21-AOAS1580

STRUCTURED HIERARCHICAL MODELS FOR PROBABILISTIC INFERENCE FROM PERTURBATION SCREENING DATA

BY SIMON DIRMEIER[a] AND NIKO BEERENWINKEL[b]

Department of Biosystems Science and Engineering, ETH Zurich, [a]*simon.dirmeier@bsse.ethz.ch,* [b]*niko.beerenwinkel@bsse.ethz.ch*

Genetic perturbation screening is an experimental method in biology to study cause and effect relationships between different biological entities. However, knocking out or knocking down genes is a highly error-prone process that complicates estimation of the effect sizes of the interventions. Here, we introduce a family of generative models, called the *structured hierarchical model* (SHM) for probabilistic inference of causal effects from perturbation screens. SHMs utilize classical hierarchical models to represent heterogeneous data and combine them with categorical Markov random fields to encode biological prior information over functionally related biological entities. The random field induces a clustering of functionally related genes which informs inference of parameters in the hierarchical model. The SHM is designed for extremely noisy data sets for which the true data generating process is difficult to model due to lack of domain knowledge or high stochasticity of the interventions. We apply the SHM to a pan-cancer genetic perturbation screen in order to identify genes that restrict the growth of an entire group of cancer cell lines and show that incorporating prior knowledge in the form of a graph improves inference of parameters.

REFERENCES

AGUIRRE, A. J., MEYERS, R. M., WEIR, B. A., VAZQUEZ, F., ZHANG, C.-Z., BEN-DAVID, U., COOK, A., HA, G., HARRINGTON, W. F. et al. (2016). Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. *Cancer Discov.* **6** 914–929.

BALDI, P., CHAUVIN, Y., HUNKAPILLER, T. and MCCLURE, M. A. (1994). Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA* **91** 1059–1063.

BETANCOURT, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. Preprint. Available at arXiv:1701.02434.

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

CHEN, M., CHO, J. and ZHAO, H. (2011). Incorporating biological pathways via a Markov random field model in genome-wide association studies. *PLoS Genet.* **7** e1001353.

CHEN, X., SHI, X., XU, X., WANG, Z., MILLS, R., LEE, C. and XU, J. (2012). A two-graph guided multitask lasso approach for eQTL mapping. In *International Connference on Artificial Intelligence and Statistics* 208–217.

COWLEY, G. S., WEIR, B. A., VAZQUEZ, F., TAMAYO, P., SCOTT, J. A., RUSIN, S., EAST-SELETSKY, A., ALI, L. D., GERATH, W. F. et al. (2014). Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci. Data* **1** 140035.

DE LUIS BALAGUER, M. A. and SOZZANI, R. (2017). Inferring gene regulatory networks in the arabidopsis root using a dynamic Bayesian network approach. In *Plant Gene Regulatory Networks* 331–348. Springer, New York.

DEPMAP, BROAD (2019). DepMap Achilles 19Q1 Public. Fileset on figshare. https://doi.org/10.6084/m9.figshare.7655150.v2

DIRMEIER, S. and BEERENWINKEL, N. (2022). Supplement to "Structured hierarchical models for probabilistic inference from perturbation screening data." https://doi.org/10.1214/21-AOAS1580SUPPA, https://doi.org/10.1214/21-AOAS1580SUPPB

Key words and phrases. Probabilistic models, hierarchical models, Markov random fields, biological network, genetic perturbation screen, interventional data, `Python`, `PyMC3`.

DIRMEIER, S., FUCHS, C., MUELLER, N. S. and THEIS, F. J. (2017). netReg: Network-regularized linear models for biological association studies. *Bioinformatics* **34** 896–898.

DIRMEIER, S., DÄCHERT, C., VAN HEMERT, M., TAS, A., OGANDO, N. S., VAN KUPPEVELD, F., BARTENSCHLAGER, R., KADERALI, L., BINDER, M. et al. (2020). Host factor prioritization for pan-viral genetic perturbation screens using random intercept models and network propagation. *PLoS Comput. Biol.* **16** 1–19.

DOENCH, J. G., HARTENIAN, E., GRAHAM, D. B., TOTHOVA, Z., HEGDE, M., SMITH, I., SULLENDER, M., EBERT, B. L., XAVIER, R. J. et al. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32** 1262–1267.

DOENCH, J. G., FUSI, N., SULLENDER, M., HEGDE, M., VAIMBERG, E. W., DONOVAN, K. F., SMITH, I., TOTHOVA, Z., WILEN, C. et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34** 184–191.

DOUDNA, J. A. and CHARPENTIER, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346** 1258096. https://doi.org/10.1126/science.1258096

DURBIN, R., EDDY, S. R., KROGH, A. and MITCHISON, G. (1998). *Biological Sequence Analysis*: *Probabilistic Models of Proteins and Nucleic Acids*. Cambridge Univ. Press, Cambridge.

EDDY, S. R. (1998). Profile hidden Markov models. *Bioinformatics* **14** 755–763.

EFRON, B. (2012). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758 https://doi.org/10.1017/CBO9780511761362

FINN, R. D., CLEMENTS, J. and EDDY, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39** W29–W37. https://doi.org/10.1093/nar/gkr367

FRIEDMAN, N., LINIAL, M., NACHMAN, I. and PE'ER, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7** 601–620.

FUSI, N., LIPPERT, C., LAWRENCE, N. D. and STEGLE, O. (2014). Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nat. Commun.* **5** 4890. https://doi.org/10.1038/ncomms5890

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

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

GELMAN, A. and SHALIZI, C. R. (2013). Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.* **66** 8–38. MR3044854 https://doi.org/10.1111/j.2044-8317.2011.02037.x

GELMAN, A., SIMPSON, D. and BETANCOURT, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy* **19** 555.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3235677

GUO, Y. and SCHUURMANS, D. (2006). Convex structure learning for Bayesian networks: Polynomial feature selection and approximate ordering. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. *UAI'*06.

HAGBERG, A., SCHULT, D. and SWART, P. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference* 11–15.

HART, T., BROWN, K. R., SIRCOULOMB, F., ROTTAPEL, R. and MOFFAT, J. (2014). Measuring error rates in genomic perturbation screens: Gold standards for human functional genomics. *Mol. Syst. Biol.* **10** 733.

HART, T., CHANDRASHEKHAR, M., AREGGER, M., STEINHART, Z., BROWN, K. R., MACLEOD, G., MIS, M., ZIMMERMANN, M., FRADET-TURCOTTE, A. et al. (2015). High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* **163** 1515–1526.

HAWE, J. S., THEIS, F. J. and HEINIG, M. (2019). Inferring interaction networks from multi-omics data. *Front. Genet.* **10** 535. https://doi.org/10.3389/fgene.2019.00535

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

IMKELLER, K., AMBROSI, G., BOUTROS, M. and HUBER, W. (2019). Modelling asymmetric count ratios in CRISPR screens to decrease experiment size and improve phenotype detection. *BioRxiv*.

JANG, E., GU, S. and POOLE, B. (2017). Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*.

JANSEN, R., YU, H., GREENBAUM, D., KLUGER, Y., KROGAN, N. J., CHUNG, S., EMILI, A., SNYDER, M., GREENBLATT, J. F. et al. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302** 449–453.

JINEK, M., CHYLINSKI, K., FONFARA, I., HAUER, M., DOUDNA, J. A. and CHARPENTIER, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337** 816–821.

KASS, R. E. and NATARAJAN, R. (2006). A default conjugate prior for variance components in generalized linear mixed models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 535–542. MR2221285 https://doi.org/10.1214/06-BA117B

KASS, R. E. and WASSERMAN, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91** 1343–1370.

KIM, S. and XING, E. P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to EQTL mapping. *Ann. Appl. Stat.* **6** 1095–1117. MR3012522 https://doi.org/10.1214/12-AOAS549

KOLLER, D. and FRIEDMAN, N. (2009). *Probabilistic Graphical Models*: *Principles and Techniques*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2778120

KORTE, A., VILHJÁLMSSON, B. J., SEGURA, V., PLATT, A., LONG, Q. and NORDBORG, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* **44** 1066–1071.

KRUMSIEK, J., SUHRE, K., ILLIG, T., ADAMSKI, J. and THEIS, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.* **5** 21. https://doi.org/10.1186/1752-0509-5-21

KUIPERS, J., THURNHERR, T., MOFFA, G., SUTER, P., BEHR, J., GOOSEN, R., CHRISTOFORI, G. and BEERENWINKEL, N. (2018). Mutational interactions define novel cancer subgroups. *Nat. Commun.* **9** 4353. https://doi.org/10.1038/s41467-018-06867-x

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

LI, Z., LI, P., KRISHNAN, A. and LIU, J. (2011). Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis. *Bioinformatics* **27** 2686–2691.

LI, W., XU, H., XIAO, T., CONG, L., LOVE, M. I., ZHANG, F., IRIZARRY, R. A., LIU, J. S., BROWN, M. et al. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15** 554.

LOH, P.-R., TUCKER, G., BULIK-SULLIVAN, B. K., VILHJALMSSON, B. J., FINUCANE, H. K., SALEM, R. M., CHASMAN, D. I., RIDKER, P. M., NEALE, B. M. et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47** 284–290.

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

MAATHUIS, M., DRTON, M., LAURITZEN, S. and WAINWRIGHT, M., eds. (2018). *Handbook of Graphical Models*. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR3889064

MARIONI, J. C., THORNE, N. P. and TAVARÉ, S. (2006). BioHMM: A heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics* **22** 1144–1146.

MEYERS, R. M., BRYAN, J. G., MCFARLAND, J. M., WEIR, B. A., SIZEMORE, A. E., XU, H., DHARIA, N. V., MONTGOMERY, P. G., COWLEY, G. S. et al. (2017). Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49** 1779–1784.

MUNOZ, D. M., CASSIANI, P. J., LI, L., BILLY, E., KORN, J. M., JONES, M. D., GOLJI, J., RUDDY, D. A., YU, K. et al. (2016). CRISPR screens provide a comprehensive assessment of cancer vulnerabilities but generate false-positive hits for highly amplified genomic regions. *Cancer Discov.* **6** 900–913.

MURPHY, K., MIAN, S. et al. (1999). Modelling gene expression data using dynamic Bayesian networks. Technical report, Computer Science Division, Univ. California.

NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 113–162. CRC Press, Boca Raton, FL. MR2858447

ONG, S. H., LI, Y., KOIKE-YUSA, H. and YUSA, K. (2017). Optimised metrics for CRISPR-KO screens with second-generation gRNA libraries. *Sci. Rep.* **7** 7384.

OUGHTRED, R., STARK, C., BREITKREUTZ, B.-J., RUST, J., BOUCHER, L., CHANG, C., KOLAS, N., O'DONNELL, L., LEUNG, G. et al. (2018). The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47** D529–D541.

PATEL, S. J., SANJANA, N. E., KISHTON, R. J., EIDIZADEH, A., VODNALA, S. K., CAM, M., GARTNER, J. J., JIA, L., STEINBERG, S. M. et al. (2017). Identification of essential genes for cancer immunotherapy. *Nature* **548** 537–542.

RAKITSCH, B., LIPPERT, C., STEGLE, O. and BORGWARDT, K. (2012). A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics* **29** 206–214.

RÄMÖ, P., DREWEK, A., ARRIEUMERLOU, C., BEERENWINKEL, N., BEN-TEKAYA, H., CARDEL, B., CASANOVA, A., CONDE-ALVAREZ, R., COSSART, P. et al. (2014). Simultaneous analysis of large-scale RNAi screens for pathogen entry. *BMC Genomics* **15** 1162.

ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, G. K. (2010). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.

SACHS, K., PEREZ, O., PE'ER, D., LAUFFENBURGER, D. A. and NOLAN, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308** 523–529.

SALVATIER, J., WIECKI, T. V. and FONNESBECK, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.* **2** e55.

SCHMICH, F., SZCZUREK, E., KREIBICH, S., DILLING, S., ANDRITSCHKE, D., CASANOVA, A., LOW, S. H., EICHER, S., MUNTWILER, S. et al. (2015). gespeR: A statistical model for deconvoluting off-target-confounded RNA interference screens. *Genome Biol.* **16** 220.

SCHUBERT, B., MADDAMSETTI, R., NYMAN, J., FARHAT, M. R. and MARKS, D. S. (2019). Genome-wide discovery of epistatic loci affecting antibiotic resistance in *Neisseria gonorrhoeae* using evolutionary couplings. *Nat. Microbiol.* **4** 328–338. https://doi.org/10.1038/s41564-018-0309-1

STANKE, M. and WAACK, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19** ii215–ii225.

SZKLARCZYK, D., GABLE, A. L., LYON, D., JUNGE, A., WYDER, S., HUERTA-CEPAS, J., SIMONOVIC, M., DONCHEVA, N. T., MORRIS, J. H. et al. (2018). STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47** D607–D613.

TSHERNIAK, A., VAZQUEZ, F., MONTGOMERY, P. G., WEIR, B. A., KRYUKOV, G., COWLEY, G. S., GILL, S., HARRINGTON, W. F., PANTEL, S. et al. (2017). Defining a cancer dependency map. *Cell* **170** 564–576.

WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

WEI, Z. and LI, H. (2007). A Markov random field model for network-based analysis of genomic data. *Bioinformatics* **23** 1537–1544.

WEI, Z. and LI, H. (2008). A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data. *Ann. Appl. Stat.* **2** 408–429. MR2415609 https://doi.org/10.1214/07--AOAS145

WU, G., FENG, X. and STEIN, L. (2010). A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.* **11** R53. https://doi.org/10.1186/gb-2010-11-5-r53

WU, X., SCOTT, D. A., KRIZ, A. J., CHIU, A. C., HSU, P. D., DADON, D. B., CHENG, A. W., TREVINO, A. E., KONERMANN, S. et al. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32** 670–676.

XU, H., XIAO, T., CHEN, C.-H., LI, W., MEYER, C. A., WU, Q., WU, D., CONG, L., ZHANG, F. et al. (2015). Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* **25** 1147–1157.

YOON, B.-J. (2009). Hidden Markov models and their applications in biological sequence analysis. *Curr. Genomics* **10** 402–415.

ZAMORA-RESENDIZ, R. and CRIVELLI, S. (2019). Structural learning of proteins using graph convolutional neural networks. *BioRxiv*.

ZHOU, X. and STEPHENS, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44** 821–824.

ZHU, S., CAO, Z., LIU, Z., HE, Y., WANG, Y., YUAN, P., LI, W., TIAN, F., BAO, Y. et al. (2019). Guide RNAs with embedded barcodes boost CRISPR-pooled screens. *Genome Biol.* **20** 20.

ZITNIK, M., AGRAWAL, M. and LESKOVEC, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34** i457–i466. https://doi.org/10.1093/bioinformatics/bty294

ZOU, M. and CONZEN, S. D. (2004). A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* **21** 71–79.

The Annals of Applied Statistics
2022, Vol. 16, No. 3, 2030–2053
https://doi.org/10.1214/21-AOAS1584

MODELING ANIMAL MOVEMENT WITH DIRECTIONAL PERSISTENCE AND ATTRACTIVE POINTS

BY GIANLUCA MASTRANTONIO[a]

Department of Mathematical Sciences, Politecnico di Torino, [a]*gianluca.mastrantonio@polito.it*

GPS technology is currently easily accessible to researchers, and many animal movement data sets are available. Two of the main features that a model which describes an animal's path can possess are directional persistence and attraction to a point in space. In this work, we propose a new approach that can have both characteristics. Our proposal is a hidden Markov model with a new emission distribution. The emission distribution models the two aforementioned characteristics, while the latent state of the hidden Markov model is needed to account for the behavioral modes. We show that the model is easy to implement in a Bayesian framework. We estimate our proposal on the motivating data that represent GPS locations of a Maremma Sheepdog recorded in Australia. The obtained results are easily interpretable and we show that our proposal outperforms the main competitive model.

REFERENCES

ABE, T. and LEY, C. (2017). A tractable, parsimonious and flexible model for cylindrical data, with applications. *Econom. Stat.* **4** 91–104. MR3702852 https://doi.org/10.1016/j.ecosta.2016.04.001

ANDERSON, C. R. and LINDZEY, F. G. (2003). Estimating cougar predation rates from gps location clusters. *J. Wildl. Manag.* **67** 307–316.

BARTON, K. A., PHILLIPS, B. L., MORALES, J. M. and TRAVIS, J. M. J. (2009). The evolution of an 'intelligent' dispersal strategy: Biased, correlated random walks in patchy landscapes. *Oikos* **118** 309–319.

BEZANSON, J., EDELMAN, A., KARPINSKI, S. and SHAH, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Rev.* **59** 65–98. MR3605826 https://doi.org/10.1137/141000671

BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 719–725.

BLACKWELL, P. (1997). Random diffusion models for animal movement. *Ecol. Model.* **100** 87–102.

BROOK, L. A., JOHNSON, C. N. and RITCHIE, E. G. (2012). Effects of predator control on behaviour of an apex predator and indirect consequences for mesopredator suppression. *J. Appl. Ecol.* **49** 1278–1286.

BROST, B. M., HOOTEN, M. B., HANKS, E. M. and SMALL, R. J. (2015). Animal movement constraints improve resource selection inference in the presence of telemetry error. *Ecology* **96** 2590–2597.

BUDERMAN, F. E., HOOTEN, M. B., IVAN, J. S. and SHENK, T. M. (2018a). Large-scale movement behavior in a reintroduced predator population. *Ecography* **41** 126–139.

BUDERMAN, F. E., HOOTEN, M. B., ALLDREDGE, M. W., HANKS, E. M. and IVAN, J. S. (2018b). Time-varying predatory behavior is primary predictor of fine-scale movement of wildland-urban cougars. *Mov. Ecol.* **6** 22.

CAGNACCI, F., BOITANI, L., POWELL, R. A. and BOYCE, M. S. (2010). Animal ecology meets gps-based radiotelemetry: A perfect storm of opportunities and challenges. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **365** 2157–2162.

CELEUX, G., FORBES, F., ROBERT, C. P. and TITTERINGTON, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Anal.* **1** 651–673. MR2282197 https://doi.org/10.1214/06-BA122

CHRIST, A., VER HOEF, J. and ZIMMERMAN, D. L. (2008). An animal movement model incorporating home range and habitat selection. *Environ. Ecol. Stat.* **15** 27–38. MR2412676 https://doi.org/10.1007/s10651-007-0036-x

CODLING, E. A. and HILL, N. A. (2005). Sampling rate effects on measurements of correlated and biased random walks. *J. Theoret. Biol.* **233** 573–588. MR2126865 https://doi.org/10.1016/j.jtbi.2004.11.008

CODLING, E. A., PLANK, M. J. and BENHAMOU, S. (2008). Random walk models in biology. *J. R. Soc. Interface* **5** 813–834.

Key words and phrases. Maremma sheepdog, step length, turning angle, CRW, BRW.

DUNN, J. E. and GIPSON, P. S. (1977). Analysis of radiotelemetry data in studies of home range. *Biometrics* **33**.

FLEMING, C. H., CALABRESE, J. M., MUELLER, T., OLSON, K. A., LEIMGRUBER, P. and FAGAN, W. F. (2014). Non-Markovian maximum likelihood estimation of autocorrelated movement processes. *Methods Ecol. Evol.* **5** 462–472.

FORTIN, D., MORALES, J. M. and BOYCE, M. S. (2005). Elk winter foraging at fine scale in Yellowstone National Park. *Oecologia* **145** 334–342.

FOX, E. B., SUDDERTH, E. B., JORDAN, M. I. and WILLSKY, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *Ann. Appl. Stat.* **5** 1020–1056. MR2840185 https://doi.org/10.1214/10-AOAS395

FRAIR, J. L., FIEBERG, J., HEBBLEWHITE, M., CAGNACCI, F., DECESARE, N. J. and PEDROTTI, L. (2010). Resolving issues of imprecise and habitat-biased locations in ecological analyses using gps telemetry data. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **365** 2187–2200.

FRIENDLY, M., MONETTE, G. and FOX, J. (2013). Elliptic insights: Understanding statistical methods through elliptical geometry. *Statist. Sci.* **28** 1–39. MR3075337 https://doi.org/10.1214/12-STS402

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

FRYXELL, J. M., HAZELL, M., BÖRGER, L., DALZIEL, B. D., HAYDON, D. T., MORALES, J. M., MCINTOSH, T. and ROSATTE, R. C. (2008). Multiple movement modes by large herbivores at multiple spatiotemporal scales. *Proc. Natl. Acad. Sci. USA* **105** 19114–19119.

GEHRING, T. M., VERCAUTEREN, K. C. and CELLAR, A. C. (2017). Good fences make good neighbors: Implementation of electric fencing for establishing effective livestock-protection dogs. *Human-Wildlife Interactions* **5** 106–111.

HANKS, E. M., HOOTEN, M. B. and ALLDREDGE, M. W. (2015). Continuous-time discrete-space models for animal movement. *Ann. Appl. Stat.* **9** 145–165. MR3341111 https://doi.org/10.1214/14-AOAS803

HARRIS, K. J. and BLACKWELL, P. G. (2013). Flexible continuous-time modelling for heterogeneous animal movement. *Ecol. Model.* **255** 29–37.

HASTIE, D. I. and GREEN, P. J. (2012). Model choice using reversible jump Markov chain Monte Carlo. *Stat. Neerl.* **66** 309–338. MR2955422 https://doi.org/10.1111/j.1467-9574.2012.00516.x

HEBBLEWHITE, M. and MERRILL, E. (2008). Modelling wildlife and uman relationships for social species with mixed-effects resource selection models. *J. Appl. Ecol.* **45** 834–844.

HOOTEN, M., JOHNSON, D., MCCLINTOCK, B. and MORALES, J. (2017). *Animal Movement*: *Statistical Models for Telemetry Data*. CRC Press.

ISHWARAN, H. and ZAREPOUR, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canad. J. Statist.* **30** 269–283. MR1926065 https://doi.org/10.2307/3315951

JAMMALAMADAKA, S. R. and KOZUBOWSKI, T. J. (2004). New families of wrapped distributions for modeling skew circular data. *Comm. Statist. Theory Methods* **33** 2059–2074. MR2103062 https://doi.org/10.1081/STA-200026570

JOHNSON, D. S., HOOTEN, M. B. and KUHN, C. E. (2013). Estimating animal resource selection from telemetry data using point process models. *J. Anim. Ecol.* **82** 1155–1164.

JOHNSON, D. S., LONDON, J. M., LEA, M.-A. and DURBAN, J. W. (2008). Continuous-time correlated random walk model for animal telemetry data. *Ecology* **89** 1208–1215.

JONSEN, I. D., FLEMMING, J. M. and MYERS, R. A. (2005). Robust state-space modeling of animal movement data. *Ecology* **86** 2874–2880.

LANGROCK, R., KING, R., MATTHIOPOULOS, J., THOMAS, L., FORTIN, D. and MORALES, J. M. (2012). Flexible and practical modeling of animal telemetry data: Hidden Markov models and extensions. *Ecology* **93** 2336–2342.

LANGROCK, R., HOPCRAFT, G., BLACKWELL, P., GOODALL, V., KING, R., NIU, M., PATTERSON, T., PEDERSEN, M., SKARIN, A. et al. (2014). Modelling group dynamic animal movement. *Methods Ecol. Evol.* **5** 190–199.

MASTRANTONIO, G. (2018). The joint projected normal and skew-normal: A distribution for poly-cylindrical data. *J. Multivariate Anal.* **165** 14–26. MR3768750 https://doi.org/10.1016/j.jmva.2017.11.006

MASTRANTONIO, G. (2022a). Supplement A to "Modeling animal movement with directional persistence and attractive points." https://doi.org/10.1214/21-AOAS1584SUPPA

MASTRANTONIO, G. (2022b). Supplement B to "'Modeling animal movement with directional persistence and attractive points." https://doi.org/10.1214/21-AOAS1584SUPPB

MASTRANTONIO, G., JONA LASINIO, G. and GELFAND, A. E. (2016). Spatio-temporal circular models with non-separable covariance structure. *TEST* **25** 331–350. MR3493522 https://doi.org/10.1007/s11749-015-0458-y

MASTRANTONIO, G., MARUOTTI, A. and JONA-LASINIO, G. (2015). Bayesian hidden Markov modelling using circular-linear general projected normal distribution. *Environmetrics* **26** 145–158. MR3324908 https://doi.org/10.1002/env.2326

MASTRANTONIO, G., GRAZIAN, C., MANCINELLI, S. and BIBBONA, E. (2019). New formulation of the logistic-Gaussian process to analyze trajectory tracking data. *Ann. Appl. Stat.* **13** 2483–2508. MR4037438 https://doi.org/10.1214/19-aoas1289

MCCLINTOCK, B. T. and MICHELOT, T. (2018). Momentuhmm: R package for generalized hidden Markov models of animal movement. *Methods Ecol. Evol.* **9** 1518–1530.

MCCLINTOCK, B. T., KING, R., THOMAS, L., MATTHIOPOULOS, J., MCCONNELL, B. J. and MORALES, J. M. (2012). A general discrete-time modeling framework for animal movement using multistate random walks. *Ecol. Monogr.* **82** 335–349.

MCCLINTOCK, B. T., JOHNSON, D. S., HOOTEN, M. B., HOEF, J. M. V. and MORALES, J. M. (2014). When to be discrete: The importance of time formulation in understanding animal movement. *Mov. Ecol.* **2** 21. https://doi.org/10.1186/s40462-014-0021-6

MCGREW, J. C. and BLAKESLEY, C. S. (1982). How komondor dogs reduce sheep losses to coyotes. *J. Range Manag.* **6** 693–696.

MERRILL, S. B. and DAVID MECH, L. (2000). Details of extensive movements by Minnesota wolves (canis lupus). *Am. Midl. Nat.* **144** 428–433.

MICHELOT, T. and BLACKWELL, P. G. (2019). State-switching continuous-time correlated random walks. *Methods Ecol. Evol.* **10** 637–649.

MICHELOT, T., LANGROCK, R. and PATTERSON, T. A. (2016). Movehmm: An R package for the statistical modelling of animal movement data using hidden Markov models. *Methods Ecol. Evol.* **7** 1308–1315.

MORALES, J. M. and ELLNER, S. P. (2002). Scaling up animal movements in heterogeneous landscapes: The importance of behavior. *Ecology* **83** 2240–2247.

MORALES, J. M., HAYDON, D. T., FRAIR, J., HOLSINGER, K. E. and FRYXELL, J. M. (2004). Extracting more out of relocation data: Building movement models as mixtures of random walks. *Ecology* **85** 2436–2445.

NATHAN, R., GETZ, W. M., REVILLA, E., HOLYOAK, M., KADMON, R., SALTZ, D. and SMOUSE, P. E. (2008). A movement ecology paradigm for unifying organismal movement research. *Proc. Natl. Acad. Sci. USA* **105** 19052–19059.

PARTON, A. and BLACKWELL, P. G. (2017). Bayesian inference for multistate 'step and turn' animal movement in continuous time. *J. Agric. Biol. Environ. Stat.* **22** 373–392. MR3692470 https://doi.org/10.1007/s13253-017-0286-5

PATTERSON, T., THOMAS, L., WILCOX, C., OVASKAINEN, O. and MATTHIOPOULOS, J. (2008). State-space models of individual animal movement. *Trends Ecol. Evol.* **23** 87–94.

PATTERSON, T. A., PARTON, A., LANGROCK, R., BLACKWELL, P. G., THOMAS, L. and KING, R. (2017). Statistical modelling of individual animal movement: An overview of key methods and a discussion of practical challenges. *AStA Adv. Stat. Anal.* **101** 399–438. MR3712406 https://doi.org/10.1007/s10182-017-0302-7

POHLE, J., LANGROCK, R., VAN BEEST, F. M. and SCHMIDT, N. M. (2017). Selecting the number of states in hidden Markov models: Pragmatic solutions illustrated using animal movement. *J. Agric. Biol. Environ. Stat.* **22** 270–293. MR3692465 https://doi.org/10.1007/s13253-017-0283-8

RIVEST, L.-P., DUCHESNE, T., NICOSIA, A. and FORTIN, D. (2016). A general angular regression model for the analysis of data on animal movement in ecology. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 445–463. MR3470586 https://doi.org/10.1111/rssc.12124

SCHULTZ, C. B. and CRONE, E. E. (2001). Edge-mediated dispersal behavior in a prairie butterfly. *Ecology* **82** 1879–1892.

VAN BOMMEL, L. and INVASIVE ANIMALS COOPERATIVE RESEARCH CENTRE (2010). *Guardian Dogs*: *Best Practice Manual for the Use of Livestock Guardian Dogs*. Invasive Animals Cooperative Research Centre.

VAN BOMMEL, L. and JOHNSON, C. N. (2012). Good dog! Using livestock guardian dogs to protect livestock from predators in Australia's extensive grazing systems. *Wildl. Res.* **39** 220–229.

VAN BOMMEL, L. and JOHNSON, C. (2014a). Data from: Where do livestock guardian dogs go? Movement patterns of free-ranging maremma sheepdogs. https://doi.org/10.5441/001/1.pv048q7v

VAN BOMMEL, L. and JOHNSON, C. N. (2014b). Where do livestock guardian dogs go? Movement patterns of free-ranging maremma sheepdogs. *PLoS ONE* **9** 1–12.

VAN BOMMEL, L. and JOHNSON, C. N. (2016). Livestock guardian dogs as surrogate top predators? How maremma sheepdogs affect a wildlife community. *Ecol. Evol.* **6** 6702–6711.

VOLANT, S., BÉRARD, C., MARTIN-MAGNIETTE, M.-L. and ROBIN, S. (2014). Hidden Markov models with mixtures as emission distributions. *Stat. Comput.* **24** 493–504. MR3223536 https://doi.org/10.1007/s11222-013-9383-7

WALTON, Z., SAMELIUS, G., ODDEN, M. and WILLEBRAND, T. (2017). Variation in home range size of red foxes vulpes vulpes along a gradient of productivity and human landscape alteration. *PLoS ONE* **12** 1–14.

WANG, F. and GELFAND, A. E. (2013). Directional data analysis under the general projected normal distribution. *Stat. Methodol.* **10** 113–127. MR2974815 https://doi.org/10.1016/j.stamet.2012.07.005