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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the
Institute of Mathematical Statistics

Articles

continued

Vol. 18, No. 3—September 2024

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*

Vol. 18, No. 3—September 2024

THE ANNALS OF APPLIED STATISTICS Vol. 18, No. 3, pp. 1773–2677 September 2024

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

***The Annals of Applied Statistics* [ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 18, Number 3, September 2024. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage pending at Cleveland, Ohio, and at additional mailing offices.

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

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 1773–1788
https://doi.org/10.1214/23-AOAS1828

MULTISCALE POISSON PROCESS APPROACHES FOR DETECTING AND ESTIMATING DIFFERENCES FROM HIGH-THROUGHPUT SEQUENCING ASSAYS

By Heejung Shim[1,a], Zhengrong Xing[2,b], Ester Pantaleo[2,c], Francesca Luca[3,d], Roger Pique-Regi[4,e] and Matthew Stephens[5,f]

[1]*School of Mathematics and Statistics and Melbourne Integrative Genomics, University of Melbourne,* [a]*heejung.shim@unimelb.edu.au*

[2]*Department of Statistics, University of Chicago,* [b]*xiamanoobix@gmail.com,* [c]*esterpantaleo@gmail.com*

[3]*Department of Obstetrics and Gynecology and Center for Molecular Medicine and Genetics, Wayne State University,* [d]*fluca@wayne.edu*

[4]*Center for Molecular Medicine and Genetics, Wayne State University,* [e]*rpique@wayne.edu*

[5]*Department of Statistics and Human Genetics, University of Chicago,* [f]*mstephens@uchicago.edu*

Estimating and testing for differences in molecular phenotypes (e.g., gene expression, chromatin accessibility, transcription factor binding) across conditions is an important part of understanding the molecular basis of gene regulation. These phenotypes are commonly measured using high-throughput sequencing assays (e.g., RNA-seq, ATAC-seq, ChIP-seq), which provide high-resolution count data that reflect how the phenotypes vary along the genome. Multiple methods have been proposed to help exploit these high-resolution measurements for differential expression analysis. However, they ignore the count nature of the data, instead using normal distributions that work well only for data with large sample sizes or high counts. Here we develop count-based methods to address this problem. We model the data for each sample using an inhomogeneous Poisson process with spatially structured underlying intensity function and then, building on multiscale models for the Poisson process, estimate and test for differences in the underlying intensity function across samples (or groups of samples). Using both simulation and real ATAC-seq data, we show that our method outperforms previous normal-based methods, especially in situations with small sample sizes or low counts.

REFERENCES

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* **129** 823–37. https://doi.org/10.1016/j.cell.2007.05.009

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S. and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132** 311–22. https://doi.org/10.1016/j.cell.2007.12.014

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10** 1213–1218. MR3211372

Busby, M. A., Stewart, C., Miller, C. A., Grzeda, K. R. and Marth, G. T. (2013). Scotty: A web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* **29** 656–657. https://doi.org/10.1093/bioinformatics/btt015

Coifman, R. R. and Donoho, D. L. (1995). Translation-invariant de-noising. In *Wavelets and Statistics* 125–150. Springer, Berlin.

Collado-Torres, L., Nellore, A., Frazee, A. C., Wilks, C., Love, M. I., Langmead, B., Irizarry, R. A., Leek, J. T. and Jaffe, A. E. (2017). Flexible expressed region analysis for RNA-seq with derfinder. *Nucleic Acids Res.* **45** e9. https://doi.org/10.1093/nar/gkw852

Key words and phrases. Multiscale Poisson processes, wavelets, differential expression analysis, high-throughput sequencing assays, high-resolution, Bayesian inference, functional data, count data, RNA-seq, DNase-seq, ATAC-seq, chromatin accessibility.

Crouse, M. S., Nowak, R. D. and Baraniuk, R. G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Process.* **46** 886–902. MR1665651 https://doi.org/10.1109/78.668544

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N. et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482** 390–4. https://doi.org/10.1038/nature10808

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464

Frazee, A. C., Sabunciyan, S., Hansen, K. D., Irizarry, R. A. and Leek, J. T. (2014). Differential expression analysis of RNA-seq data at single-base resolution. *Biostatistics* **15** 413–426. https://doi.org/10.1093/biostatistics/kxt053

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S. et al. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6** 283–9. https://doi.org/10.1038/nmeth.1313

Johnson, D. S., Mortazavi, A., Myers, R. M. and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316** 1497–502. https://doi.org/10.1126/science.1141319

Kolaczyk, E. D. (1999). Bayesian multiscale models for Poisson processes. *J. Amer. Statist. Assoc.* **94** 920–933. MR1723303 https://doi.org/10.2307/2670007

Law, C. W., Chen, Y., Shi, W. and Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15** 1.

Lee, W. and Morris, J. S. (2016). Identification of differentially methylated loci using wavelet-based functional mixed models. *Bioinformatics* **32** 664–672. https://doi.org/10.1093/bioinformatics/btv659

Liu, Y., Zhou, J. and White, K. P. (2014). RNA-seq differential expression studies: More sequence or more replication? *Bioinformatics* **30** 301–304. https://doi.org/10.1093/bioinformatics/btt688

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.

Luca, F., Maranville, J. C., Richards, A. L., Witonsky, D. B., Stephens, M. and Rienzo, A. D. (2013). Genetic, functional and molecular features of glucocorticoid receptor binding. *PLoS ONE* **8** e61654. https://doi.org/10.1371/journal.pone.0061654

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

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18** 1509–17. https://doi.org/10.1101/gr.079558.108

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K. et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448** 553–60. https://doi.org/10.1038/nature06008

Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A. and Coombes, K. R. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics* **64** 479–489, 667. MR2432418 https://doi.org/10.1111/j.1541-0420.2007.00895.x

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5** 621–8. https://doi.org/10.1038/nmeth.1226

Moyerbrailean, G. A., Davis, G. O., Harvey, C. T., Watza, D., Wen, X., Pique-Regi, R. and Luca, F. (2015). A high-throughput RNA-seq approach to profile transcriptional responses. *Sci. Rep.* **5** 14976. https://doi.org/10.1038/srep14976

Pique-Regi, R., Degner, J. F., Pai, A. A., Boyle, A. P., Song, L., Lee, B.-K., Gaffney, D. J., Gilad, Y. and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21** 447–55. https://doi.org/10.1101/gr.112623.110

Robinson, D. G. and Storey, J. D. (2014). subSeq: Determining appropriate sequencing depth through efficient read subsampling. *Bioinformatics* **30** 3424–3426. https://doi.org/10.1093/bioinformatics/btu552

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. https://doi.org/10.1093/bioinformatics/btp616

Shim, H. and Stephens, M. (2015). Wavelet-based genetic association analysis of functional phenotypes arising from high-throughput sequencing assays. *Ann. Appl. Stat.* **9** 665–686. MR3371330 https://doi.org/10.1214/14-AOAS776

Shim, H., Xing, Z., Pantaleo, E., Luca, F., Pique-Regi, R. and Stephens, M. (2024). Supplement to "Multiscale Poisson process approaches for detecting and estimating differences from high-throughput sequencing assays." https://doi.org/10.1214/23-AOAS1828SUPPA, https://doi.org/10.1214/23-AOAS1828SUPPB, https://doi.org/10.1214/23-AOAS1828SUPPC

SMYTH, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3** Art. 3, 29. MR2101454 https://doi.org/10.2202/1544-6115.1027

STEPHENS, M. (2017). False discovery rates: A new deal. *Biostatistics* **18** 275–294. MR3824755 https://doi.org/10.1093/biostatistics/kxw041

STOREY, J. D., BASS, A. J., DABNEY, A. and ROBINSON, D. (2020). qvalue: Q-value estimation for false discovery rate control R package version 2.20.0.

TARAZONA, S., GARCÍA-ALCALDE, F., DOPAZO, J., FERRER, A. and CONESA, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Res.* **21** 2213–2223.

TIMMERMANN, K. E. and NOWAK, R. D. (1999). Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging. *IEEE Trans. Inf. Theory* **45** 846–862. MR1682515 https://doi.org/10.1109/18.761328

WAKEFIELD, J. (2009). Bayes factors for genome-wide association studies: Comparison with P-values. *Genet. Epidemiol.* **33** 79–86. https://doi.org/10.1002/gepi.20359

WANG, E. T., SANDBERG, R., LUO, S., KHREBTUKOVA, I., ZHANG, L., MAYR, C., KINGSMORE, S. F., SCHROTH, G. P. and BURGE, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* **456** 470–6. https://doi.org/10.1038/nature07509

XING, Z., CARBONETTO, P. and STEPHENS, M. (2021). Flexible signal denoising via flexible empirical Bayes shrinkage. *J. Mach. Learn. Res.* **22** Paper No. 93, 28. MR4279744

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 1789–1814
https://doi.org/10.1214/23-AOAS1845

DYNAMIC MODELING AND ONLINE MONITORING OF TENSOR DATA STREAMS WITH APPLICATION TO PASSENGER FLOW SURVEILLANCE

BY YIFAN LI[1,a], CHUNJIE WU[2,b], WENDONG LI[3,c], FUGEE TSUNG[4,d] AND JIANHUA GUO[5,e]

[1]*School of Statistics and Data Science, Nanjing Audit University,* [a]*lyfmgr@163.com*
[2]*School of Statistics and Management, Shanghai University of Finance and Economics,* [b]*wumaths@mail.shufe.edu.cn*
[3]*KLATASDS-MOE, School of Statistics, East China Normal University,* [c]*wendongli01@gmail.com*
[4]*Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology,* [d]*season@ust.hk*
[5]*School of Mathematics and Statistics, Beijing Technology and Business University,* [e]*jhguo@btbu.edu.cn*

Passenger flow surveillance in urban transport systems has emerged as a major global issue for smart city management. Governments are taking proper measures to monitor passenger flow in order to maintain social stability and to prevent unexpected group events. It is critical to develop a passenger flow surveillance system that continuously monitors the passenger flow over time and triggers a signal as soon as the passenger flow begins to deteriorate so that timely government intervention can be implemented. In this paper passenger flow surveillance is novelly formulated as dynamic modeling and online monitoring of tensor data streams. Existing tensor monitoring methods either rely heavily on the assumption that the tensor coefficients exhibit a low-rank structure or are inapplicable to general-order tensors. We propose a unified monitoring framework based on the tensor normal distribution to overcome these challenges. We begin by developing a tensor model selection procedure that ensures that the chosen tensor structure strikes a balance between model complexity and estimation accuracy. Then we propose an online estimation procedure to dynamically estimate the tensor parameters on which sequential change-detection procedures, using the generalized likelihood ratio test, are proposed. Extensive simulations and an analysis of real passenger flow data in Hong Kong demonstrate the efficacy of our approach.

REFERENCES

CHATTERJEE, S. and QIU, P. (2009). Distribution-free cumulative sum control charts using bootstrap-based control limits. *Ann. Appl. Stat.* **3** 349–369. MR2668711 https://doi.org/10.1214/08-AOAS197

CHEN, R., YANG, D. and ZHANG, C.-H. (2022). Factor models for high-dimensional tensor time series. *J. Amer. Statist. Assoc.* **117** 94–116. MR4399070 https://doi.org/10.1080/01621459.2021.1912757

COLOSIMO, B. M. and GRASSO, M. (2018). Spatially weighted PCA for monitoring video image data with application to additive manufacturing. *J. Qual. Technol.* **50** 391–417.

DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963 https://doi.org/10.1093/biomet/68.1.265

GAO, X., SHEN, W., ZHANG, L., HU, J., FORTIN, N. J., FROSTIG, R. D. and OMBAO, H. (2021). Regularized matrix data clustering and its application to image analysis. *Biometrics* **77** 890–902. MR4320665 https://doi.org/10.1111/biom.13354

GÓMEZ, A. M. E., LI, D. and PAYNABAR, K. (2022). An adaptive sampling strategy for online monitoring and diagnosis of high-dimensional streaming data. *Technometrics* **64** 253–269. MR4410918 https://doi.org/10.1080/00401706.2021.1967198

GREENEWALD, K., ZHOU, S. and HERO, A. III (2019). Tensor graphical lasso (TeraLasso). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 901–931. MR4025402 https://doi.org/10.1111/rssb.12339

HAN, D. and TSUNG, F. (2006). A reference-free Cuscore chart for dynamic mean change detection and a unified framework for charting performance comparison. *J. Amer. Statist. Assoc.* **101** 368–386. MR2268053 https://doi.org/10.1198/016214505000000556

Key words and phrases. Passenger flow surveillance, online monitoring, tensor data streams, tensor normal distribution, model selection, control chart.

HAN, F., ZHAO, T. and LIU, H. (2013). CODA: High dimensional copula discriminant analysis. *J. Mach. Learn. Res.* **14** 629–671. MR3033343

HAWKINS, D. M. and MABOUDOU-TCHAO, E. M. (2008). Multivariate exponentially weighted moving covariance matrix. *Technometrics* **50** 155–166. MR2439876 https://doi.org/10.1198/004017008000000163

HE, S., YIN, J., LI, H. and WANG, X. (2014). Graphical model selection and estimation for high dimensional tensor data. *J. Multivariate Anal.* **128** 165–185. MR3199836 https://doi.org/10.1016/j.jmva.2014.03.007

HOFF, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Anal.* **6** 179–196. MR2806238 https://doi.org/10.1214/11-BA606

HOFF, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *Ann. Appl. Stat.* **9** 1169–1193. MR3418719 https://doi.org/10.1214/15-AOAS839

JIA, Q., ZHANG, Y. and CHEN, W. (2018). Image-based process monitoring using projective nonnegative-tensor factorization. *IEEE Trans. Ind. Electron.*. https://doi.org/10.1109/TIE.2018.2833027

KHANZADEH, M., TIAN, W., YADOLLAHI, A., DOUDE, H. R., TSCHOPP, M. A. and BIAN, L. (2018). Dual process monitoring of metal-based additive manufacturing using tensor decomposition of thermal image streams. *Addit. Manuf.* **23** 443–456.

KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056 https://doi.org/10.1137/07070111X

LEE, W., MCCORMICK, T. H., NEIL, J., SODJA, C. and CUI, Y. (2022). Anomaly detection in large-scale networks with latent space models. *Technometrics* **64** 241–252. MR4410917 https://doi.org/10.1080/00401706.2021.1952900

LI, C. and ZHANG, H. (2021). Tensor quantile regression with application to association between neuroimages and human intelligence. *Ann. Appl. Stat.* **15** 1455–1477. MR4316657 https://doi.org/10.1214/21-aoas1475

LI, L. and ZHANG, X. (2017). Parsimonious tensor response regression. *J. Amer. Statist. Assoc.* **112** 1131–1146. MR3735365 https://doi.org/10.1080/01621459.2016.1193022

LI, W., TSUNG, F., SONG, Z., ZHANG, K. and XIANG, D. (2021). Multi-sensor based landslide monitoring via transfer learning. *J. Qual. Technol.* **53** 474–487.

LI, W., XIANG, D., TSUNG, F. and PU, X. (2020a). A diagnostic procedure for high-dimensional data streams via missed discovery rate control. *Technometrics* **62** 84–100. MR4058601 https://doi.org/10.1080/00401706.2019.1575284

LI, Y., WU, C., LI, W., FUGEE, T. and GUO, J. (2024). Supplement to "Dynamic modeling and online monitoring of tensor data streams with application to passenger flow surveillance." https://doi.org/10.1214/23-AOAS1845SUPPA, https://doi.org/10.1214/23-AOAS1845SUPPB

LI, Z., YAN, H., ZHANG, C. and TSUNG, F. (2020b). Long-short term spatiotemporal tensor prediction for passenger flow profile. *IEEE Robot. Autom. Lett.* **5** 5010–5017.

MAI, Q., ZHANG, X., PAN, Y. and DENG, K. (2022). A doubly enhanced EM algorithm for model-based tensor clustering. *J. Amer. Statist. Assoc.* **117** 2120–2134. MR4528493 https://doi.org/10.1080/01621459.2021.1904959

MAI, Q. and ZOU, H. (2015). Sparse semiparametric discriminant analysis. *J. Multivariate Anal.* **135** 175–188. MR3306434 https://doi.org/10.1016/j.jmva.2014.12.009

MANCEUR, A. M. and DUTILLEUL, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *J. Comput. Appl. Math.* **239** 37–49. MR2991957 https://doi.org/10.1016/j.cam.2012.09.017

MEI, Y. (2010). Efficient scalable schemes for monitoring a large number of data streams. *Biometrika* **97** 419–433. MR2650748 https://doi.org/10.1093/biomet/asq010

PAN, Y., MAI, Q. and ZHANG, X. (2019). Covariate-adjusted tensor classification in high dimensions. *J. Amer. Statist. Assoc.* **114** 1305–1319. MR4011781 https://doi.org/10.1080/01621459.2018.1497500

QIU, P. (2014). *Introduction to Statistical Process Control*. CRC Press/CRC, Boca Raton, FL.

QIU, P., LI, W. and LI, J. (2020). A new process control chart for monitoring short-range serially correlated data. *Technometrics* **62** 71–83. MR4058600 https://doi.org/10.1080/00401706.2018.1562988

REN, H., ZOU, C., CHEN, N. and LI, R. (2022). Large-scale datastreams surveillance via pattern-oriented-sampling. *J. Amer. Statist. Assoc.* **117** 794–808. MR4436313 https://doi.org/10.1080/01621459.2020.1819295

SUN, W. W. and LI, L. (2019). Dynamic tensor clustering. *J. Amer. Statist. Assoc.* **114** 1894–1907. MR4047308 https://doi.org/10.1080/01621459.2018.1527701

VEERAVALLI, V. V. (2001). Decentralized quickest change detection. *IEEE Trans. Inf. Theory* **47** 1657–1665. MR1830119 https://doi.org/10.1109/18.923755

XIE, Y. and SIEGMUND, D. (2013). Control chart for dynamic process monitoring with an application to air pollution surveillance. In 2013 *Information Theory and Applications Workshop* (*ITA*) 2013 1–20.

YAN, H., PAYNABAR, K. and SHI, J. (2014). Image-based process monitoring using low-rank tensor decomposition. *IEEE Trans. Autom. Sci. Eng.* **12** 216–227.

YAN, H., PAYNABAR, K. and SHI, J. (2017). Anomaly detection in images with smooth background via smooth-sparse decomposition. *Technometrics* **59** 102–114. MR3604193 https://doi.org/10.1080/00401706.2015.1102764

YAN, H., PAYNABAR, K. and SHI, J. (2018). Real-time monitoring of high-dimensional functional data streams via spatio-temporal smooth sparse decomposition. *Technometrics* **60** 181–197. MR3804247 https://doi.org/10.1080/00401706.2017.1346522

ZHANG, C., CHEN, N. and WU, J. (2020). Spatial rank-based high-dimensional monitoring through random projection. *J. Qual. Technol.* **52** 111–127.

ZOU, C. and QIU, P. (2009). Multivariate statistical process control using LASSO. *J. Amer. Statist. Assoc.* **104** 1586–1596. MR2750580 https://doi.org/10.1198/jasa.2009.tm08128

ZOU, C., TSUNG, F. and LIU, Y. (2008). A change point approach for phase I analysis in multistage processes. *Technometrics* **50** 344–356. MR2528657 https://doi.org/10.1198/004017008000000307

ZOU, C., WANG, Z., JIANG, W. and ZI, X. (2015). An efficient online monitoring method for high-dimensional data streams. *Technometrics* **57** 374–387. MR3384952 https://doi.org/10.1080/00401706.2014.940089

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 1815–1839
https://doi.org/10.1214/23-AOAS1858

CONTINUOUS AND ATLAS-FREE ANALYSIS OF BRAIN STRUCTURAL CONNECTIVITY

BY WILLIAM CONSAGRA[1,a], MARTIN COLE[2,b], XING QIU[2,c] AND ZHENGWU ZHANG[3,d]

[1]*Psychiatry Neuroimaging Laboratory, Harvard Medical School,* [a]*wconsagra@bwh.harvard.edu*
[2]*Department of Biostatistics and Computational Biology, University of Rochester Medical Center,* [b]*martin_cole@urmc.rochester.edu,* [c]*xing_qiu@urmc.rochester.edu*
[3]*Department of Statistics and Operations Research, University of North Carolina at Chapel Hill,* [d]*zhengwu_zhang@unc.edu*

Brain structural networks are often represented as discrete adjacency matrices with elements summarizing the connectivity between pairs of regions of interest (ROIs). These ROIs are typically determined a priori using a brain atlas. The choice of atlas is often arbitrary and can lead to a loss of important connectivity information at the sub-ROI level. This work introduces an atlas-free framework that overcomes these issues by modeling brain connectivity using smooth random functions. In particular, we assume that the observed pattern of white matter fiber tract endpoints is driven by a latent random function defined over a product manifold domain. To facilitate statistical analysis of these high-dimensional functional data objects, we develop a novel algorithm to construct a data-driven reduced-rank function space that offers a desirable trade-off between computational complexity and flexibility. Using real data from the Human Connectome Project, we show that our method outperforms state-of-the-art approaches that use the traditional atlas-based structural connectivity representation on a variety of connectivity analysis tasks. We further demonstrate how our method can be used to detect localized regions and connectivity patterns associated with group differences.

REFERENCES

ALLEN, G. (2012). Sparse higher-order principal components analysis. In *AISTATS*.

AMBROSEN, K. S., ESKILDSEN, S. F., HINNE, M., KRUG, K., LUNDELL, H., SCHMIDT, M. N., VAN GERVEN, M. A., MØRUP, M. and DYRBY, T. B. (2020). Validation of structural brain connectivity networks: The impact of scanning parameters. *NeuroImage* **204** 116207.

ARROYO, J., ATHREYA, A., CAPE, J., CHEN, G., PRIEBE, C. E. and VOGELSTEIN, J. T. (2021). Inference for multiple heterogeneous networks with a common invariant subspace. *J. Mach. Learn. Res.* **22** 142. MR4318498

BASSER, P., MATTIELLO, J. and LEBIHAN, D. (1994). Estimation of the effective self-diffusion tensor from the NMR spin echo. *J. Magn. Reson., Ser. B* **103** 247–254.

BASSER, P. J., PAJEVIC, S., PIERPAOLI, C., DUDA, J. and ALDROUBI, A. (2000). In vivo fiber tractography using dt-mri data. *Magn. Reson. Med.* **44** 625–632.

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

BEZDEK, J. C. and HATHAWAY, R. J. (2003). Convergence of alternating optimization. *Neural Parallel Sci. Comput.* **11** 351–368. MR2020725

BOROVITSKIY, V., AZANGULOV, I., TERENIN, A., MOSTOWSKY, P., DEISENROTH, M. P. and DURRANDE, N. (2021). Matern Gaussian processes on graphs. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

BOUZAS, P. R., VALDERRAMA, M. J., AGUILERA, A. M. and RUIZ-FUENTES, N. (2006). Modelling the mean of a doubly stochastic Poisson process by functional data analysis. *Comput. Statist. Data Anal.* **50** 2655–2667. MR2227341 https://doi.org/10.1016/j.csda.2005.04.015

CAI, Y., FANG, G. and LI, P. (2021). A note on sparse generalized eigenvalue problem. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Key words and phrases. Point process, functional data analysis, structural connectivity, neuroimaging, high dimensional.

CHEN, K., DELICADO, P. and MÜLLER, H.-G. (2017). Modelling function-valued stochastic processes, with applications to fertility dynamics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 177–196. MR3597969 https://doi.org/10.1111/rssb.12160

CHUNG, J., BRIDGEFORD, E., ARROYO, J., PEDIGO, B. D., SAAD-ELDIN, A., GOPALAKRISHNAN, V., XIANG, L., PRIEBE, C. E. and VOGELSTEIN, J. T. (2021). Statistical connectomics. *Annu. Rev. Stat. Appl.* **8** 463–492. MR4243556 https://doi.org/10.1146/annurev-statistics-042720-023234

CHUNG, M. (2006). Heat kernel smoothing on unit sphere. In *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro*, 2006 992–995.

COLE, M., MURRAY, K., ST-ONGE, E., RISK, B., ZHONG, J., SCHIFITTO, G., DESCOTEAUX, M. and ZHANG, Z. (2021). Surface-based connectivity integration: An atlas-free approach to jointly study functional and structural connectivity. *Hum. Brain Mapp.* **42** 3481–3499.

CONSAGRA, W., COLE, M., QIU, X. and ZHANG, Z. (2024). Supplement to "Continuous and atlas-free analysis of brain structural connectivity." https://doi.org/10.1214/23-AOAS1858SUPPA, https://doi.org/10.1214/23-AOAS1858SUPPB

CONSAGRA, W., COLE, M. and ZHANG, Z. (2022). Analyzing brain structural connectivity as continuous random functions. In *Medical Image Computing and Computer Assisted Intervention* 276–285. Springer, Cham.

CONSAGRA, W., VENKATARAMAN, A. and QIU, X. (2023). Efficient multidimensional functional data analysis using marginal product basis systems. *J. Comput. Graph. Statist.* **0** 1–11.

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

DESTRIEUX, C., FISCHL, B., DALE, A. and HALGREN, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage* **53** 1–15.

DURANTE, D., DUNSON, D. B. and VOGELSTEIN, J. T. (2017). Nonparametric Bayes modeling of populations of networks. *J. Amer. Statist. Assoc.* **112** 1516–1530. MR3750873 https://doi.org/10.1080/01621459.2016.1219260

ELIAS, L. J., BULMAN-FLEMING, M. and MCMANUS, I. (1999). Visual temporal asymmetries are related to asymmetries in linguistic perception. *Neuropsychologia* **37** 1243–1249.

ESTLE, S. J., GREEN, L., MYERSON, J. and HOLT, D. D. (2006). Differential effects of amount on temporal and probability discounting of gains and losses. *Mem. Cogn.* **34** 914–928.

FISCHL, B., SERENO, M. I. and DALE, A. M. (1999). Cortical surface-based analysis: Ii: Inflation, flattening, and a surface-based coordinate system. *NeuroImage* **9** 195–207.

FORNITO, A., ZALESKY, A. and BREAKSPEAR, M. (2013). Graph analysis of the human connectome: Promise, progress, and pitfalls. *NeuroImage* **80** 426–444. Mapping the Connectome.

GERSHON, R. C., WAGSTER, M. V., HENDRIE, H. C., FOX, N. A., COOK, K. F. and NOWINSKI, C. J. (2013). Nih toolbox for assessment of neurological and behavioral function. *Neurology* **80** S2–S6.

GIRARD, G., WHITTINGSTALL, K., DERICHE, R. and DESCOTEAUX, M. (2014). Towards quantitative connectivity analysis: Reducing tractography biases. *NeuroImage* **98** 266–278.

GLASSER, M. F., COALSON, T. S., ROBINSON, E. C., HACKER, C. D., HARWELL, J., YACOUB, E., UGURBIL, K., ANDERSSON, J., BECKMANN, C. F. et al. (2016). A multi-modal parcellation of human cerebral cortex. *Nature* **536** 171–178.

GLASSER, M. F., SOTIROPOULOS, S. N., WILSON, J. A., COALSON, T. S., FISCHL, B., ANDERSSON, J. L., XU, J., JBABDI, S., WEBSTER, M. et al. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80** 105–124.

GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

GUTMAN, B., LEONARDO, C., JAHANSHAD, N., HIBAR, D., ESCHENBURG, K., NIR, T., VILLALON, J. and THOMPSON, P. (2014). Registering cortical surfaces based on whole-brain structural connectivity and continuous connectivity analysis. *Med. Image Comput. Comput. Assist. Interv.* **17** 161–168.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6** 65–70. MR0538597

HOU, K. and SO, A. M.-C. (2014). Hardness and approximation results for L_p-ball constrained homogeneous polynomial optimization problems. *Math. Oper. Res.* **39** 1084–1108. MR3279759 https://doi.org/10.1287/moor.2014.0644

JUNG, S., AHN, J. and JEON, Y. (2019). Penalized orthogonal iteration for sparse estimation of generalized eigenvalue problem. *J. Comput. Graph. Statist.* **28** 710–721. MR4007752 https://doi.org/10.1080/10618600.2019.1568014

LILA, E., ASTON, J. A. D. and SANGALLI, L. M. (2016). Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *Ann. Appl. Stat.* **10** 1854–1879. MR3592040 https://doi.org/10.1214/16-AOAS975

LYNCH, B. and CHEN, K. (2018). A test of weak separability for multi-way functional data, with application to brain connectivity studies. *Biometrika* **105** 815–831. MR3877867 https://doi.org/10.1093/biomet/asy048

MANSOUR, S., SEGUIN, C., SMITH, R. E. and ZALESKY, A. (2022). Connectome spatial smoothing (css): Concepts, methods, and evaluation. *NeuroImage* **250** 118930.

MEINSHAUSEN, N., MAATHUIS, M. H. and BÜHLMANN, P. (2011). Asymptotic optimality of the Westfall–Young permutation procedure for multiple testing under dependence. *Ann. Statist.* **39** 3369–3391. MR3012412 https://doi.org/10.1214/11-AOS946

MOYER, D., GUTMAN, B. A., FASKOWITZ, J., JAHANSHAD, N. and THOMPSON, P. M. (2017). Continuous representations of brain connectivity using spatial point processes. *Med. Image Anal.* **41** 32–39.

NIELSEN, A. M. and WITTEN, D. (2018). The multiple random dot product graph model. Statistics Methodology. arXiv.

ODUM, A. L. (2011). Delay discounting: Trait variable? *Behav. Process.* **87** 1–9.

OLSON, E. A., COLLINS, P. F., HOOPER, C. J., MUETZEL, R., LIM, K. O. and LUCIANA, M. (2009). White matter integrity predicts delay discounting behavior in 9- to 23-year-olds: A diffusion tensor imaging study. *J. Cogn. Neurosci.* **21** 1406–1421.

OWENS, M. M., GRAY, J. C., AMLUNG, M. T., OSHRI, A., SWEET, L. H. and MACKILLOP, J. (2017). Neuroanatomical foundations of delayed reward discounting decision making. *NeuroImage* **161** 261–270.

PANARETOS, V. M. and ZEMEL, Y. (2016). Amplitude and phase variation of point processes. *Ann. Statist.* **44** 771–812. MR3476617 https://doi.org/10.1214/15-AOS1387

PARK, H.-J. and FRISTON, K. (2013). Structural and functional brain networks: From connections to cognition. *Science* **342**.

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

PINI, A. and VANTINI, S. (2016). The interval testing procedure: A general framework for inference in functional data analysis. *Biometrics* **72** 835–845. MR3545676 https://doi.org/10.1111/biom.12476

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*. Springer, New York. https://doi.org/10.1007/b98888

ROGERS, R. D., OWEN, A. M., MIDDLETON, H. C., WILLIAMS, E. J., PICKARD, J. D., SAHAKIAN, B. J. and ROBBINS, T. W. (1999). Choosing between small, likely rewards and large, unlikely rewards activates inferior and orbital prefrontal cortex. *J. Neurosci.* **19** 9029–9038.

SCHAEFER, A., KONG, R., GORDON, E. M., LAUMANN, T. O., ZUO, X. N., HOLMES, A. J., EICKHOFF, S. B. and YEO, B. T. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* **28** 3095–3114.

SCHUMAKER, L. L. (2015). *Spline Functions—Computational Methods*. SIAM, Philadelphia, PA. MR3430816 https://doi.org/10.1137/1.9781611973907.ch1

SILVERMAN, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Statist.* **24** 1–24. MR1389877 https://doi.org/10.1214/aos/1033066196

SMITH, S. M., JENKINSON, M., WOOLRICH, M. W., BECKMANN, C. F., BEHRENS, T. E., JOHANSEN-BERG, H., BANNISTER, P. R., DE LUCA, M., DROBNJAK, I. et al. (2004). Advances in functional and structural mr image analysis and implementation as fsl. *NeuroImage* **23** S208–S219.

SRIVASTAVA, A., KLASSEN, E., JOSHI, S. H. and JERMYN, I. H. (2011). Shape analysis of elastic curves in Euclidean spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 1415–1428.

ST-ONGE, E., DADUCCI, A., GIRARD, G. and DESCOTEAUX, M. (2018). Surface-enhanced tractography. *NeuroImage* **169** 524–539.

STANGER, C., ELTON, A., RYAN, S. R., JAMES, G. A., BUDNEY, A. J. and KILTS, C. D. (2013). Neuroeconomics and adolescent substance abuse: Individual differences in neural networks and delay discounting. *J. Amer. Acad. Child Adolesc. Psych.* **52** 747–755.e6.

TAN, K. M., WANG, Z., LIU, H. and ZHANG, T. (2018). Sparse generalized eigenvalue problem: Optimal statistical rates via truncated Rayleigh flow. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 1057–1086. MR3874310 https://doi.org/10.1111/rssb.12291

TOURNIER, J. D., CALAMANTE, F. and CONNELLY, A. (2007). Robust determination of the fibre orientation distribution in diffusion MRI: Non-negativity constrained super-resolved spherical deconvolution. *NeuroImage* **35** 1459–1472.

TUCH, D. S. (2004). Q-ball imaging. *Magn. Reson. Med.* **52** 1358–1372.

WANG, L., ZHANG, Z. and DUNSON, D. (2019). Common and individual structure of brain networks. *Ann. Appl. Stat.* **13** 85–112. MR3937422 https://doi.org/10.1214/18-AOAS1193

WANG, Q., CHEN, C., CAI, Y., LI, S., ZHAO, X., ZHENG, L., ZHANG, H., LIU, J., CHEN, C. et al. (2016). Dissociated neural substrates underlying impulsive choice and impulsive action. *NeuroImage* **134** 540–549.

WANG, S., ARROYO, J., VOGELSTEIN, J. T. and PRIEBE, C. E. (2021). Joint embedding of graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **43** 1324–1336.

WROBEL, J., ZIPUNNIKOV, V., SCHRACK, J. and GOLDSMITH, J. (2019). Registration for exponential family functional data. *Biometrics* **75** 48–57. MR3953706 https://doi.org/10.1111/biom.12963

WU, L., QIU, X., YUAN, Y. and WU, H. (2019). Parameter estimation and variable selection for big systems of linear ordinary differential equations: A matrix-based approach. *J. Amer. Statist. Assoc.* **114** 657–667. MR3963170 https://doi.org/10.1080/01621459.2017.1423074

WU, S., MÜLLER, H.-G. and ZHANG, Z. (2013). Functional data analysis for point processes with rare events. *Statist. Sinica* **23** 1–23. MR3076156

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106 https://doi.org/10.1214/009053605000000660

YUAN, X.-T. and ZHANG, T. (2013). Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.* **14** 899–925. MR3063614

ZALESKY, A., FORNITO, A. and BULLMORE, E. T. (2010). Network-based statistic: Identifying differences in brain networks. *NeuroImage* **53** 1197–1207.

ZALESKY, A., FORNITO, A., HARDING, I. H., COCCHI, L., YÜCEL, M., PANTELIS, C. and BULLMORE, E. T. (2010). Whole-brain anatomical networks: Does the choice of nodes matter? *NeuroImage* **50** 970–983.

ZHANG, Z., ALLEN, G. I., ZHU, H. and DUNSON, D. (2019). Tensor network factorizations: Relationships between brain structural connectomes and traits. *NeuroImage* **197** 330–343.

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 1840–1857
https://doi.org/10.1214/23-AOAS1859
© Institute of Mathematical Statistics, 2024

A BOOTSTRAP MODEL COMPARISON TEST FOR IDENTIFYING GENES WITH CONTEXT-SPECIFIC PATTERNS OF GENETIC REGULATION

BY MYKHAYLO M. MALAKHOV[1,a] BEN DAI[2,c], XIAOTONG T. SHEN[3,d] AND WEI PAN[1,b]

[1]*Division of Biostatistics and Health Data Science, University of Minnesota,* [a]*malak039@umn.edu,* [b]*panxx014@umn.edu*
[2]*Department of Statistics, The Chinese University of Hong Kong,* [c]*bendai@cuhk.edu.hk*
[3]*School of Statistics, University of Minnesota,* [d]*xshen@umn.edu*

Understanding how genetic variation affects gene expression is essential for a complete picture of the functional pathways that give rise to complex traits. Although numerous studies have established that many genes are differentially expressed in distinct human tissues and cell types, no tools exist for identifying the genes whose expression is differentially regulated. Here we introduce DRAB (differential regulation analysis by bootstrapping), a gene-based method for testing whether patterns of genetic regulation are significantly different between tissues or other biological contexts. DRAB first leverages the elastic net to learn context-specific models of local genetic regulation and then applies a novel bootstrap-based model comparison test to check their equivalency. Unlike previous model comparison tests, our proposed approach can determine whether population-level models have equal predictive performance by accounting for the variability of feature selection and model training. We validated DRAB on mRNA expression data from a variety of human tissues in the Genotype-Tissue Expression (GTEx) Project. DRAB yielded biologically reasonable results and had sufficient power to detect genes with tissue-specific regulatory profiles while effectively controlling false positives. By providing a framework that facilitates the prioritization of differentially regulated genes, our study enables future discoveries on the genetic architecture of molecular phenotypes.

REFERENCES

ALLMAN, J. M., TETREAULT, N. A., HAKEEM, A. Y., MANAYE, K. F., SEMENDEFERI, K., ERWIN, J. M., PARK, S., GOUBERT, V. and HOF, P. R. (2011). The von Economo neurons in the frontoinsular and anterior cingulate cortex. *Ann. N.Y. Acad. Sci.* **1225** 59–71. https://doi.org/10.1111/j.1749-6632.2011.06011.x

BARBEIRA, A. N., BONAZZOLA, R., GAMAZON, E. R., LIANG, Y., PARK, Y., KIM-HELLMUTH, S., WANG, G., JIANG, Z., ZHOU, D. et al. (2021). Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22** 49. https://doi.org/10.1186/s13059-020-02252-4

BARBEIRA, A. N., DICKINSON, S. P., BONAZZOLA, R., ZHENG, J., WHEELER, H., TORRES, J. M., TORSTENSON, E. S., SHAH, K. P., GARCIA, T. et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9** 1825. https://doi.org/10.1038/s41467-018-03621-1

BARTHAS, F., SELLMEIJER, J., HUGEL, S., WALTISPERGER, E., BARROT, M. and YALCIN, I. (2015). The anterior cingulate cortex is a critical hub for pain-induced depression. *Biol. Psychiatry* **77** 236–245. https://doi.org/10.1016/j.biopsych.2014.08.004

BAUR, B., SHIN, J., ZHANG, S. and ROY, S. (2020). Data integration for inferring context-specific gene regulatory networks. *Curr. Opin. Syst. Biol.* **23** 38–46. https://doi.org/10.1016/j.coisb.2020.09.005

BEASLEY, C. L., PENNINGTON, K., BEHAN, A., WAIT, R., DUNN, M. J. and COTTER, D. (2006). Proteomic analysis of the anterior cingulate cortex in the major psychiatric disorders: Evidence for disease-associated changes. *Proteomics* **6** 3414–3425. https://doi.org/10.1002/pmic.200500069

BHUVA, D. D., CURSONS, J., SMYTH, G. K. and DAVIS, M. J. (2019). Differential co-expression-based detection of conditional relationships in transcriptional data: Comparative analysis and application to breast cancer. *Genome Biol.* **20** 236. https://doi.org/10.1186/s13059-019-1851-8

Key words and phrases. Differential regulation, gene expression, model selection test, penalized regression, transcriptome prediction.

Bullard, J. H., Purdom, E., Hansen, K. D. and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinform.* **11** 94. https://doi.org/10.1186/1471-2105-11-94

Cadiou, S. and Slama, R. (2021). Instability of variable-selection algorithms used to identify true predictors of an outcome in intermediate-dimension epidemiologic studies. *Epidemiology* **32** 402–411. https://doi.org/10.1097/EDE.0000000000001340

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M. and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4** 7. https://doi.org/10.1186/s13742-015-0047-8

Dai, B., Shen, X. and Pan, W. (2022). Significance tests of feature relevance for a black-box learner. *IEEE Trans. Neural Netw. Learn. Syst.* https://doi.org/10.1109/TNNLS.2022.3185742

Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software `hdi`. *Statist. Sci.* **30** 533–558. MR3432840 https://doi.org/10.1214/15-STS527

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903 https://doi.org/10.1201/9780429246593

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**. https://doi.org/10.18637/jss.v033.i01

Fryett, J. J., Morris, A. P. and Cordell, H. J. (2020). Investigation of prediction accuracy and the impact of sample size, ancestry, and tissue in transcriptome-wide association studies. *Genet. Epidemiol.* **44** 425–441. https://doi.org/10.1002/gepi.22290

Gallagher, M. D. and Chen-Plotkin, A. S. (2018). The post-GWAS era: From association to function. *Am. J. Hum. Genet.* **102** 717–730. https://doi.org/10.1016/j.ajhg.2018.04.002

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., GTEx Consortium et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47** 1091–1098. https://doi.org/10.1038/ng.3367

Gefen, T., Papastefan, S. T., Rezvanian, A., Bigio, E. H., Weintraub, S., Rogalski, E., Mesulam, M. M. and Geula, C. (2018). Von Economo neurons of the anterior cingulate across the lifespan and in Alzheimer's disease. *Cortex* **99** 69–77. https://doi.org/10.1016/j.cortex.2017.10.015

Gillies, C. E., Putler, R., Menon, R., Otto, E., Yasutake, K., Nair, V., Hoover, P., Lieb, D., Li, S. et al. (2018). An eQTL landscape of kidney tissue in human nephrotic syndrome. *Am. J. Hum. Genet.* **103** 232–244. https://doi.org/10.1016/j.ajhg.2018.07.004

Graffelman, J. and Moreno, V. (2013). The mid p-value in exact tests for Hardy–Weinberg equilibrium. *Stat. Appl. Genet. Mol. Biol.* **12** 433–448. MR3101039 https://doi.org/10.1515/sagmb-2012-0039

Grueneberg, A. and de los Campos, G. (2019). BGData—a suite of R packages for genomic analysis with big data. *G3 Genes|Genomes|Genetics* **9** 1377–1383. https://doi.org/10.1534/g3.119.400018

Guo, X., Lin, W., Wen, W., Huyghe, J., Bien, S., Cai, Q., Harrison, T., Chen, Z., Qu, C. et al. (2021). Identifying novel susceptibility genes for colorectal cancer risk from a transcriptome-wide association study of 125,478 subjects. *Gastroenterology* **160** 1164–1178. https://doi.org/10.1053/j.gastro.2020.08.062

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J. C., Boomsma, D. I. et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48** 245–252. https://doi.org/10.1038/ng.3506

He, R., Xue, H., Pan, W. and For the Alzheimer's Disease Neuroimaging Initiative (2022). Statistical power of transcriptome-wide association studies. *Genet. Epidemiol.* **46** 572–588. https://doi.org/10.1002/gepi.22491

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., Yu, Z., Li, B., Gu, J. et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet.* **51** 568–576. https://doi.org/10.1038/s41588-019-0345-7

Jia, J. and Yu, B. (2010). On model selection consistency of the elastic net when $p \gg n$. *Statist. Sinica* **20** 595–611. MR2682632

Kalousis, A., Prados, J. and Hilario, M. (2007). Stability of feature selection algorithms: A study on high-dimensional spaces. *Knowl. Inf. Syst.* **12** 95–116. https://doi.org/10.1007/s10115-006-0040-8

Keys, K. L., Mak, A. C. Y., White, M. J., Eckalbar, W. L., Dahl, A. W., Mefford, J., Mikhaylova, A. V., Contreras, M. G., Elhawary, J. R. et al. (2020). On the cross-population generalizability of gene expression prediction models. *PLoS Genet.* **16** e1008927. https://doi.org/10.1371/journal.pgen.1008927

LIN, Z., XUE, H., MALAKHOV, M. M., KNUTSON, K. A. and PAN, W. (2022). Accounting for nonlinear effects of gene expression identifies additional associated genes in transcriptome-wide association studies. *Hum. Mol. Genet.* **31** 2462–2470. https://doi.org/10.1093/hmg/ddac015

LONSDALE, J., THOMAS, J., SALVATORE, M., PHILLIPS, R., LO, E., SHAD, S., HASZ, R., WALTERS, G., GARCIA, F. et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45** 580–585. https://doi.org/10.1038/ng.2653

MALAKHOV, M. M., DAI, B., SHEN, X. T. and PAN, W. (2024). Supplement to "A bootstrap model comparison test for identifying genes with context-specific patterns of genetic regulation." https://doi.org/10.1214/23-AOAS1859SUPPA, https://doi.org/10.1214/23-AOAS1859SUPPB, https://doi.org/10.1214/23-AOAS1859SUPPC

NOGUEIRA, S., SECHIDIS, K. and BROWN, G. (2018). On the stability of feature selection algorithms. *J. Mach. Learn. Res.* **18** 174. MR3827062

OKORO, P. C., SCHUBERT, R., GUO, X., JOHNSON, W. C., ROTTER, J. I., HOESCHELE, I., LIU, Y., IM, H. K., LUKE, A. et al. (2021). Transcriptome prediction performance across machine learning models and diverse ancestries. *Hum. Genet. Genomics Adv.* **2** 100019. https://doi.org/10.1016/j.xhgg.2020.100019

PAN, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18** 546–554. https://doi.org/10.1093/bioinformatics/18.4.546

POLITIS, D. N. and ROMANO, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22** 2031–2050. MR1329181 https://doi.org/10.1214/aos/1176325770

PURCELL, S. and CHANG, C. (2023). PLINK 1.90. Version beta 7 (16 Jan 2023). https://www.cog-genomics.org/plink/1.9

SEYEDNASROLLAH, F., LAIHO, A. and ELO, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.* **16** 59–70. https://doi.org/10.1093/bib/bbt086

SONESON, C. and DELORENZI, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform.* **14** 91. https://doi.org/10.1186/1471-2105-14-91

THE GTEX CONSORTIUM (2020). The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* **369** 1318–1330. https://doi.org/10.1126/science.aaz1776

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242 https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

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

WASSERMAN, L., RAMDAS, A. and BALAKRISHNAN, S. (2020). Universal inference. *Proc. Natl. Acad. Sci. USA* **117** 16880–16890. MR4242731 https://doi.org/10.1073/pnas.1922664117

WHEELER, H. E., SHAH, K. P., BRENNER, J., GARCIA, T., AQUINO-MICHAELS, K., GTEX CONSORTIUM, COX, N. J., NICOLAE, D. L. and IM, H. K. (2016). Survey of the heritability and sparse architecture of gene expression traits across human tissues. *PLoS Genet.* **12** e1006423. https://doi.org/10.1371/journal.pgen.1006423

WIGGINTON, J. E., CUTLER, D. J. and ABECASIS, G. R. (2005). A note on exact tests of Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.* **76** 887–893. https://doi.org/10.1086/429864

WONG, A. K., SEALFON, R. S. G., THEESFELD, C. L. and TROYANSKAYA, O. G. (2021). Decoding disease: From genomes to networks to phenotypes. *Nat. Rev. Genet.* **22** 774–790. https://doi.org/10.1038/s41576-021-00389-x

WU, L., SHI, W., LONG, J., GUO, X., MICHAILIDOU, K., BEESLEY, J., BOLLA, M. K., SHU, X.-O., LU, Y. et al. (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.* **50** 968–978. https://doi.org/10.1038/s41588-018-0132-x

YANG, T., WU, C., WEI, P. and PAN, W. (2020). Integrating DNA sequencing and transcriptomic data for association analyses of low-frequency variants and lipid traits. *Hum. Mol. Genet.* **29** 515–526. https://doi.org/10.1093/hmg/ddz314

YAZDANI, A., MENDEZ-GIRALDEZ, R., YAZDANI, A., KOSOROK, M. R. and ROUSSOS, P. (2020). Differential gene regulatory pattern in the human brain from schizophrenia using transcriptomic-causal network. *BMC Bioinform.* **21** 469. https://doi.org/10.1186/s12859-020-03753-6

YUAN, M. and LIN, Y. (2007). On the non-negative garrote estimator. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 143–161. MR2325269 https://doi.org/10.1111/j.1467-9868.2007.00581.x

ZHAO, S., WITTEN, D. and SHOJAIE, A. (2021). In defense of the indefensible: A very naïve approach to high-dimensional inference. *Statist. Sci.* **36** 562–577. MR4323053 https://doi.org/10.1214/20-sts815

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

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 1858–1878
https://doi.org/10.1214/23-AOAS1860
© Institute of Mathematical Statistics, 2024

PATIENT RECRUITMENT USING ELECTRONIC HEALTH RECORDS UNDER SELECTION BIAS: A TWO-PHASE SAMPLING FRAMEWORK

BY GUANGHAO ZHANG[1,a], LAUREN J. BEESLEY[2,d], BHRAMAR MUKHERJEE[1,b] AND XU SHI[1,c]

[1]*Department of Biostatistics, University of Michigan,* [a]*ghzhang@umich.edu,* [b]*bhramar@umich.edu,* [c]*shixu@umich.edu*
[2]*Statistical Sciences Group, Los Alamos National Laboratory,* [d]*lvandervort@lanl.gov*

Electronic health records (EHRs) are increasingly recognized as a cost-effective resource for patient recruitment in clinical research. However, how to optimally select a cohort from millions of individuals to answer a scientific question of interest remains unclear. Consider a study to estimate the mean or mean difference of an expensive outcome. Inexpensive auxiliary covariates predictive of the outcome may often be available in patients' health records, presenting an opportunity to recruit patients selectively, which may improve efficiency in downstream analyses. In this paper we propose a two-phase sampling design that leverages available information on auxiliary covariates in EHR data. A key challenge in using EHR data for multiphase sampling is the potential selection bias, because EHR data are not necessarily representative of the target population. Extending existing literature on two-phase sampling design, we derive an optimal two-phase sampling method that improves efficiency over random sampling while accounting for the potential selection bias in EHR data. We demonstrate the efficiency gain from our sampling design via simulation studies and an application evaluating the prevalence of hypertension among U.S. adults leveraging data from the Michigan Genomics Initiative, a longitudinal biorepository in Michigan Medicine.

REFERENCES

BARRETT, J. E., CAKIROGLU, A., BUNCE, C., SHAH, A. and DENAXAS, S. (2020). Selective recruitment designs for improving observational studies using electronic health records. *Stat. Med.* **39** 2556–2567. MR4119749 https://doi.org/10.1002/sim.8556

BEAULIEU-JONES, B. K., GREENE, C. S. et al. (2016). Semi-supervised learning of the electronic health record for phenotype stratification. *J. Biomed. Inform.* **64** 168–178.

BEESLEY, L. J. and MUKHERJEE, B. (2022). Statistical inference for association studies using electronic health records: Handling both selection bias and outcome misclassification. *Biometrics* **78** 214–226. MR4408582 https://doi.org/10.1111/biom.13400

BEESLEY, L. J., SALVATORE, M., FRITSCHE, L. G., PANDIT, A., RAO, A., BRUMMETT, C., WILLER, C. J., LISABETH, L. D. and MUKHERJEE, B. (2020). The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. *Stat. Med.* **39** 773–800. MR4067765 https://doi.org/10.1002/sim.8445

BENNETT, M., VIELMA, J. P. and ZUBIZARRETA, J. R. (2020). Building representative matched samples with multi-valued treatments in large observational studies. *J. Comput. Graph. Statist.* **29** 744–757. MR4191240 https://doi.org/10.1080/10618600.2020.1753532

BOWER, J. K., BOLLINGER, C. E., FORAKER, R. E., HOOD, D. B., SHOBEN, A. B. and LAI, A. M. (2017). Active use of electronic health records (EHRs) and personal health records (PHRs) for epidemiologic research: Sample representativeness and nonresponse bias in a study of women during pregnancy. *eGEMs* **5** 1263.

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

BROWN, C. D., HIGGINS, M., DONATO, K. A., ROHDE, F. C., GARRISON, R., OBARZANEK, E. et al. (2000). Body mass index and the prevalence of hypertension and dyslipidemia. *Obes. Res.* **8** 605–619.

Key words and phrases. Auxiliary information, electronic health records, selection bias, study design, two-phase sampling.

CHANG, W.-T., WENG, S.-F., HSU, C.-H., SHIH, J.-Y., WANG, J.-J., WU, C.-Y. and CHEN, Z.-C. (2016). Prognostic factors in patients with pulmonary hypertension—a nationwide cohort study. *J. Amer. Heart Assoc.* **5**.

COWIE, M. R., BLOMSTER, J. I., CURTIS, L. H., DUCLAUX, S., FORD, I., FRITZ, F. et al. (2017). Electronic health records to facilitate clinical research. *Clin. Res. Cardiol.* **106** 1–9.

EFFOE, V. S., KATULA, J. A., KIRK, J. K., PEDLEY, C. F., BOLLHALTER, L. Y., BROWN, W. M. et al. (2016). The use of electronic medical records for recruitment in clinical trials: Findings from the lifestyle intervention for treatment of diabetes trial. *Trials* **17** 496.

ELLIOT, R. M. (2013). Combining data from probability and non-probability samples using pseudo-weights. *Surv. Pract.*

ESPINHEIRA, P. and SILVA, A. D. O. (2018). Nonlinear simplex regression models. Preprint. Available at arXiv:1805.10843.

FOX, B. D., AZOULAY, L., DELL'ANIELLO, S., LANGLEBEN, D., LAPI, F., BENISTY, J. and SUISSA, S. (2014). The use of antidepressants and the risk of idiopathic pulmonary arterial hypertension. *Can. J. Cardiol.* **30** 1633–1639. https://doi.org/10.1016/j.cjca.2014.09.031

GILBERT, P. B., YU, X. and ROTNITZKY, A. (2014). Optimal auxiliary-covariate-based two-phase sampling design for semiparametric efficient estimation of a mean or mean difference, with application to clinical trials. *Stat. Med.* **33** 901–917. MR3249030 https://doi.org/10.1002/sim.6006

GOLDSTEIN, B. A., BHAVSAR, N. A., PHELAN, M. and PENCINA, M. J. (2016). Controlling for informed presence bias due to the number of health encounters in an electronic health record. *Amer. J. Epidemiol.* **184** 847–855.

HANEUSE, S. and DANIELS, M. (2016). A general framework for considering selection bias in EHR-based studies: What data are observed and why? *eGEMs* **4** 16.

HÄYRINEN, K., SARANTO, K. and NYKÄNEN, P. (2008). Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *Int. J. Med. Inform.* **77** 291–304. https://doi.org/10.1016/j.ijmedinf.2007.09.001

HEMKENS, L. G., CONTOPOULOS-IOANNIDIS, D. G. and IOANNIDIS, J. P. (2016). Routinely collected data and comparative effectiveness evidence: Promises and limitations. *CMAJ*, *Can. Med. Assoc. J.* **188** E158–E164.

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* **15** 199–236.

JOYCE, E., WANG, S., MOTAMED, M., KIDWELL, K. M. and HENRY, N. L. (2021). Associations between preexisting nociplastic pain and early discontinuation of aromatase inhibitor therapy in breast cancer. *J. Clin. Oncol.* **39** 12068–12068.

KIESCHNICK, R. and MCCULLOUGH, B. D. (2003). Regression analysis of variates observed on $(0, 1)$: Percentages, proportions and fractions. *Stat. Model.* **3** 193–213. MR2005473 https://doi.org/10.1191/1471082X03st053oa

LEVIS, A. W., MUKHERJEE, R., WANG, R. and HANEUSE, S. (2022). Double sampling and semiparametric methods for informatively missing data. Preprint. Available at arXiv:2204.02432.

MCCORD, K. A. and HEMKENS, L. G. (2019). Using electronic health records for clinical trials: Where do we stand and where can we go? *CMAJ*, *Can. Med. Assoc. J.* **191** E128–E133.

MCISAAC, M. A. and COOK, R. J. (2015). Adaptive sampling in two-phase designs: A biomarker study for progression in arthritis. *Stat. Med.* **34** 2899–2912. MR3375988 https://doi.org/10.1002/sim.6523

PHELAN, M., BHAVSAR, N. A. and GOLDSTEIN, B. A. (2017). Illustrating informed presence bias in electronic health records data: How patient interactions with a health system can impact inference. *eGEMs* **5** 22.

PINTO, E. (2007). Blood pressure and ageing. *Postgrad. Med. J.* **83** 109–114. https://doi.org/10.1136/pgmj.2006.048371

ROTNITZKY, A. and ROBINS, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* **82** 805–820. MR1380816 https://doi.org/10.1093/biomet/82.4.805

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

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

SAHAY, B., NGUYEN, C. Q. and YAMAMOTO, J. K. (2017). Conserved HIV epitopes for an effective HIV vaccine. *J. Clin. Cell. Immunol.* **8**.

SCHREIWEIS, B., TRINCZEK, B., KÖPCKE, F., LEUSCH, T., MAJEED, R. W., WENK, J., BERGH, B., OHMANN, C., RÖHRIG, R. et al. (2014). Comparison of electronic health record system functionalities to support the patient recruitment process in clinical trials. *Int. J. Med. Inform.* **83** 860–868. https://doi.org/10.1016/j.ijmedinf.2014.08.005

SHI, X., PAN, Z. and MIAO, W. (2023). Data integration in causal inference. *Wiley Interdiscip. Rev.: Comput. Stat.* **15** Paper No. e1581, 17. MR4544393 https://doi.org/10.1002/wics.1581

SHORTREED, S. M., COOK, A. J., COLEY, R. Y., BOBB, J. F. and NELSON, J. C. (2019). Challenges and opportunities for using big health care data to advance medical science and public health. *Amer. J. Epidemiol.* **188** 851–861.

SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. MR1092986

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

THADANI, S. R., WENG, C., BIGGER, J. T., ENNEVER, J. F. and WAJNGURT, D. (2009). Electronic screening improves efficiency in clinical trial recruitment. *J. Amer. Med. Inform. Assoc.* **16** 869–873. https://doi.org/10.1197/jamia.M3119

TRIPEPI, G., JAGER, K. J., DEKKER, F. W. and ZOCCALI, C. (2010). Selection bias and information bias in clinical research. *Nephron, Clin. Pract.* **115** c94–c99. https://doi.org/10.1159/000312871

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

WU, H., TOTI, G., MORLEY, K. I., IBRAHIM, Z., FOLARIN, A., KARTOGLU, I., JACKSON, R., AGRAWAL, A., STRINGER, C. et al. (2017). SemEHR: Surfacing semantic data from clinical notes in electronic health records for tailored care, trial recruitment, and clinical research. *Lancet* **390** S97.

WU, K.-H. H., HORNSBY, W. E., KLUNDER, B., KRAUSE, A., DRISCOLL, A., KULKA, J., BICKETT-HICKOK, R., FELLOWS, A., GRAHAM, S. et al. (2021). Exposure and risk factors for COVID-19 and the impact of staying home on Michigan residents. *PLoS ONE* **16** 0246447.

ZHANG, G., BEESLEY, L. J, MUKHERJEE, B. and SHI, X. (2024). Supplement to “Patient recruitment using electronic health records under selection bias: A two-phase sampling framework.” https://doi.org/10.1214/23-AOAS1860SUPPA, https://doi.org/10.1214/23-AOAS1860SUPPB

ZHANG, P., QIU, Z., PENG, Z. and ZENGUO, Q. (2014). Regression analysis of proportional data using simplex distribution. *Sci. China Math.* (*Chinese Version*) **44** 89–104.

ZHANG, Y., LIU, M., NEYKOV, M. and CAI, T. (2022). Prior adaptive semi-supervised learning with application to EHR phenotyping. *J. Mach. Learn. Res.* **23** Paper No. [83], 25. MR4576668

ZOLLA-PAZNER, S. (2004). Identifying epitopes of HIV-1 that induce protective antibodies. *Nat. Rev., Immunol.* **4** 199–210.

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 1879–1898
https://doi.org/10.1214/23-AOAS1862
© Institute of Mathematical Statistics, 2024

A NOVEL BAYESIAN MODEL FOR ASSESSING INTRATUMOR HETEROGENEITY OF TUMOR INFILTRATING LEUKOCYTES WITH MULTIREGION GENE EXPRESSION SEQUENCING

BY PENG YANG[1,a], SHAWNA M. HUBERT[3,e], P. ANDREW FUTREAL[4,g], XINGZHI SONG[4,h], JIANHUA ZHANG[4,i], J. JACK LEE[2,b], IGNACIO WISTUBA[5,j], YING YUAN[2,c], JIANJUN ZHANG[3,f], AND ZIYI LI[2,d]

[1]*Department of Statistics, Rice University,* [a]*py11@rice.edu*

[2]*Department of Biostatistics, The University of Texas MD Anderson Cancer Center,* [b]*jjlee@mdanderson.org,* [c]*yyuan@mdanderson.org,* [d]*zli16@mdanderson.org*

[3]*Department of Thoracic Head Neck Medical Oncology, The University of Texas MD Anderson Cancer Center,* [e]*smhubert@mdanderson.org,* [f]*jzhang20@mdanderson.org*

[4]*Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center,* [g]*afutreal@mdanderson.org,* [h]*xsong3@mdanderson.org,* [i]*jzhang22@mdanderson.org*

[5]*Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center,* [j]*iiwistuba@mdanderson.org*

Intratumor heterogeneity (ITH) of tumor-infiltrated leukocytes (TILs) is an important phenomenon of cancer biology with potentially profound clinical impacts. Multiregion gene expression sequencing data provide a promising opportunity that allows for explorations of TILs and their intratumor heterogeneity for each subject. Although several existing methods are available to infer the proportions of TILs, considerable methodological gaps exist for evaluating intratumor heterogeneity of TILs with multiregion gene expression data. Here we develop ICeITH, immune cell estimation reveals intratumor heterogeneity, a Bayesian hierarchical model that borrows cell-type profiles as prior knowledge to decompose mixed bulk data while accounting for the within-subject correlations among tumor samples. ICeITH quantifies intratumor heterogeneity by the variability of targeted cellular compositions. Through extensive simulation studies, we demonstrate that ICeITH is more accurate in measuring relative cellular abundance and evaluating intratumor heterogeneity compared with existing methods. We also assess the ability of ICeITH to stratify patients by their intratumor heterogeneity score and associate the estimations with the survival outcomes. Finally, we apply ICeITH to two multiregion gene expression datasets from lung cancer studies to classify patients into different risk groups according to the ITH estimations of targeted TILs that shape either pro- or antitumor processes. In conclusion, ICeITH is a useful tool to evaluate intratumor heterogeneity of TILs from multiregion gene expression data.

REFERENCES

ABDULJABBAR, K., RAZA, S., ROSENTHAL, R., JAMAL-HANJANI, M., VEERIAH, S., AKARCA, A., LUND, T., MOORE, D., SALGADO, R. et al. (2020). Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat. Med.* **26** 1054–1062.

ANDRADE BARBOSA, B., ASTEN, S., OH, J., FARINA-SARASQUETA, A., VERHEIJ, J., DIJK, F., LAARHOVEN, H., YLSTRA, B., GARCIA VALLEJO, J. et al. (2021). Bayesian log-normal deconvolution for enhanced in silico microdissection of bulk gene expression data. *Nat. Commun.* **12** 1–13.

BACHER, R. and KENDZIORSKI, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **17** 1–14.

BINDEA, G., MLECNIK, B., TOSOLINI, M., KIRILOVSKY, A., WALDNER, M., OBENAUF, A., ANGELL, H., FREDRIKSEN, T., LAFONTAINE, L. et al. (2013). Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39** 782–795.

Key words and phrases. Bayesian hierarchical model, deconvolution, intratumor heterogeneity, RNA-seq data, TRACERx.

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

CHEN, C., LEUNG, Y. Y., IONITA, M., WANG, L.-S. and LI, M. (2022). Omnibus and robust deconvolution scheme for bulk RNA sequencing data integrating multiple single-cell reference sets and prior biological knowledge. *Bioinformatics* **38** 4530–4536. https://doi.org/10.1093/bioinformatics/btac563

CHEN, G., NING, B. and SHI, T. (2019). Single-cell RNA-seq technologies and related computational data analysis. *Front. Genet.* 317.

DE SOUSA, V. M. L. and CARVALHO, L. (2018). Heterogeneity in lung cancer. *Pathobiology* **85** 96–107. https://doi.org/10.1159/000487440

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. and GINGERAS, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29** 15–21. https://doi.org/10.1093/bioinformatics/bts635

DONG, M., THENNAVAN, A., URRUTIA, E., LI, Y., PEROU, C. M., ZOU, F. and JIANG, Y. (2021). SCDC: Bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief. Bioinform.* **22** 416–427. https://doi.org/10.1093/bib/bbz166

FAVERO, F., JOSHI, T., MARQUARD, A. M., BIRKBAK, N. J., KRZYSTANEK, M., LI, Q., SZALLASI, Z. and EKLUND, A. C. (2015). Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26** 64–70. https://doi.org/10.1093/annonc/mdu479

FENTON, L. (1960). The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions On Communications Systems* **8** 57–67.

GAUDREAU, P., NEGRAO, M., MITCHELL, K., REUBEN, A., CORSINI, E., LI, J., KARPINETS, T., WANG, Q., DIAO, L. et al. (2021). Neoadjuvant chemotherapy increases cytotoxic T cell, tissue resident memory T cell, and B cell infiltration in resectable NSCLC. *J. Thorac. Oncol.* **16** 127–139.

HUBBARD, T., BARKER, D., BIRNEY, E., CAMERON, G., CHEN, Y., CLARK, L., COX, T., CUFF, J., CURWEN, V. et al. (2002). The ensembl genome database project. *Nucleic Acids Res.* **30** 38–41.

HWANG, B., LEE, J. and BANG, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50** 1–14.

JAMAL-HANJANI, M., WILSON, G. A., MCGRANAHAN, N., BIRKBAK, N. J., WATKINS, T. B. K., VEERIAH, S., SHAFI, S., JOHNSON, D. H., MITTER, R. et al. (2017). Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376** 2109–2121. https://doi.org/10.1056/NEJMoa1616288

JANISZEWSKA, M. (2020). The microcosmos of intratumor heterogeneity: The space-time of cancer evolution. *Oncogene* **39** 2031–2039. https://doi.org/10.1038/s41388-019-1127-5

JIA, Q., WU, W., WANG, Y., ALEXANDER, P., SUN, C., GONG, Z., CHENG, J., SUN, H., GUAN, Y. et al. (2018). Local mutational diversity drives intratumoral immune heterogeneity in non-small cell lung cancer. *Nat. Commun.* **9** 1–10.

LI, B. and DEWEY, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12** 323. https://doi.org/10.1186/1471-2105-12-323

LI, B., SEVERSON, E., PIGNON, J., ZHAO, H., LI, T., NOVAK, J., JIANG, P., SHEN, H., ASTER, J. et al. (2016). Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biol.* **17** 1–16.

LINSLEY, P. S., SPEAKE, C., WHALEN, E. and CHAUSSABEL, D. (2014). Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis. *PLoS ONE* **9** e109760. https://doi.org/10.1371/journal.pone.0109760

MEISTER, M., BELOUSOV, A., XU, E., SCHNABEL, P., WARTH, A. and HOOFMANN, H. et al. (2014). Intratumor heterogeneity of gene expression profiles in early stage non-small cell lung cancer. *J Bioinf Res Stud.* **1** 1.

NEWMAN, A. M., LIU, C. L., GREEN, M. R., GENTLES, A. J., FENG, W., XU, Y., HOANG, C. D., DIEHN, M. and ALIZADEH, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12** 453–457. https://doi.org/10.1038/nmeth.3337

POTTER, S. (2018). Single-cell RNA sequencing for the study of development, physiology and disease. *Nature Reviews Nephrology* **14** 479–492.

RACLE, J. and GFELLER, D. (2020). EPIC: A tool to estimate the proportions of different cell types from bulk gene expression data. *Bioinformatics For Cancer Immunotherapy* 233–248.

RACLE, J., JONGE, K., BAUMGAERTNER, P., SPEISER, D. and GFELLER, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6** e26476.

REUBEN, A., GITTELMAN, R., GAO, J., ZHANG, J., YUSKO, E., WU, C., EMERSON, R., ZHANG, J., TIPTON, C. et al. (2017). TCR repertoire intratumor heterogeneity in localized lung adenocarcinomas: An association with predicted neoantigen heterogeneity and postsurgical RecurrenceTCR intratumor heterogeneity and relapse in lung cancer. *Cancer Discov.* **7** 1088–1097.

ROBINSON, M. and OSHLACK, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11** 1–9.

ROONEY, M. S., SHUKLA, S. A., WU, C. J., GETZ, G. and HACOHEN, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160** 48–61. https://doi.org/10.1016/j.cell.2014.12.033

SALIBA, A.-E., WESTERMANN, A. J., GORSKI, S. A. and VOGEL, J. (2014). Single-cell RNA-seq: Advances and future challenges. *Nucleic Acids Res.* **42** 8845–8860. https://doi.org/10.1093/nar/gku555

SASAKI, Y. et al. (2007). The truth of the F-measure. *Teach Tutor Mater.* **1** 1–5.

SATO, S., SANJO, H., TAKEDA, K., NINOMIYA-TSUJI, J., YAMAMOTO, M., KAWAI, T., MATSUMOTO, K., TAKEUCHI, O. and AKIRA, S. (2005). Essential function for the kinase TAK1 in innate and adaptive immune responses. *Nat. Immunol.* **6** 1087–1095. https://doi.org/10.1038/ni1255

SCHOLKOPF, B., SMOLA, A. J., WILLIAMSON, R. C. and BARTLETT, P. L. (2000). New support vector algorithms. *Neural Comput.* **12** 1207–1245. https://doi.org/10.1162/089976600300015565

TEH, Y., KURIHARA, K. and WELLING, M. (2007). Collapsed variational inference for HDP. *Adv. Neural Inf. Process. Syst.* **20**.

TEH, Y., NEWMAN, D. and WELLING, M. (2006). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Adv. Neural Inf. Process. Syst.* **19**.

WANG, P. and BLUNSOM, P. (2013). Collapsed variational Bayesian inference for hidden Markov models. *Artificial Intelligence And Statistics* 599–607.

WHITESIDE, T., VUJANOVIC, N. and HERBERMAN, R. (1998). Natural killer cells and tumor therapy. *Specificity, Function, And Development Of Nk Cells* 221–244.

WILSON, D. R., JIN, C., IBRAHIM, J. G. and SUN, W. (2020). ICeD-T provides accurate estimates of immune cell abundance in tumor samples by allowing for aberrant gene expression patterns. *J. Amer. Statist. Assoc.* **115** 1055–1065. MR4143449 https://doi.org/10.1080/01621459.2019.1654874

YANG, P. (2022). Immune cell-type estimation reveals intratumor heterogeneity. *GitHub*.

YANG, P., HUBERT, S. M, FUTREAL, P. A, SONG, X., ZHANG, J., LEE, J. J, WISTUBA, I., YUAN, Y., ZHANG, J. and LI, Z. (2024). Supplement to “A novel Bayesian model for assessing intratumor heterogeneity of tumor infiltrating leukocytes with multiregion gene expression sequencing.” https://doi.org/10.1214/23-AOAS1862SUPPA, https://doi.org/10.1214/23-AOAS1862SUPPB

ZHANG, J., FUJIMOTO, J., ZHANG, J., WEDGE, D., SONG, X., ZHANG, J., SETH, S., CHOW, C., CAO, Y. et al. (2014). Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346** 256–259.

ZHANG, Y., PARMIGIANI, G. and JOHNSON, W. E. (2020). *NAR Genomics Bioinform.* **2** lqaa078. https://doi.org/10.1093/nargab/lqaa078

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 1899–1923
https://doi.org/10.1214/23-AOAS1863

SCALABLE MULTIPLE NETWORK INFERENCE WITH THE JOINT GRAPHICAL HORSESHOE

BY CAMILLA LINGJÆRDE[1,a], BENJAMIN P. FAIRFAX[2,d], SYLVIA RICHARDSON[1,b] AND HÉLÈNE RUFFIEUX[1,c]

[1]*MRC Biostatistics Unit, University of Cambridge,* [a]*camilla.lingjaerde@mrc-bsu.cam.ac.uk,* [b]*sylvia.richardson@mrc-bsu.cam.ac.uk,* [c]*helene.ruffieux@mrc-bsu.cam.ac.uk*

[2]*Department of Oncology, MRC Weatherall Institute for Molecular Medicine, University of Oxford,* [d]*benjamin.fairfax@oncology.ox.ac.uk*

Network models are useful tools for modelling complex associations. In statistical omics such models are increasingly popular for identifying and assessing functional relationships and pathways. If a Gaussian graphical model is assumed, conditional independence is determined by the nonzero entries of the inverse covariance (precision) matrix of the data. The Bayesian graphical horseshoe estimator provides a robust and flexible framework for precision matrix inference, as it introduces local, edge-specific parameters which prevent over-shrinkage of nonzero off-diagonal elements. However, its applicability is currently limited in statistical omics settings, which often involve high-dimensional data from multiple conditions that might share common structures. We propose: (i) a scalable expectation conditional maximisation (ECM) algorithm for the original graphical horseshoe and (ii) a novel joint graphical horseshoe estimator, which borrows information across multiple related networks to improve estimation. We show numerically that our single-network ECM approach is more scalable than the existing graphical horseshoe Gibbs implementation, while achieving the same level of accuracy. We also show that our joint-network proposal successfully leverages shared edge-specific information between networks while still retaining differences, outperforming state-of-the-art methods at any level of network similarity. Finally, we leverage our approach to clarify gene regulation activity within and across immune stimulation conditions in monocytes, and formulate hypotheses on the pathogenesis of immune-mediated diseases.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) 267–281. Akad. Kiadó, Budapest. MR0483125

AKIRAV, E. M., RUDDLE, N. H. and HEROLD, K. C. (2011). The role of AIRE in human autoimmune disease. *Nat. Rev. Endocrinol.* **7** 25–33. https://doi.org/10.1038/nrendo.2010.200

BAKER, L. A., ALLIS, C. D. and WANG, G. G. (2008). PHD fingers in human diseases: Disorders arising from misinterpreting epigenetic marks. *Mutat. Res.* **647** 3–12.

BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Anal.* **12** 1105–1131. MR3724980 https://doi.org/10.1214/16-BA1028

BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. (2019). Lasso meets horseshoe: A survey. *Statist. Sci.* **34** 405–427. MR4017521 https://doi.org/10.1214/19-STS700

BISWAS, S. K. and MANTOVANI, A. (2010). Macrophage plasticity and interaction with lymphocyte subsets: Cancer as a paradigm. *Nat. Immunol.* **11** 889–896. https://doi.org/10.1038/ni.1937

BUSATTO, C. and STINGO, F. C. (2023). Inference of multiple high-dimensional networks with the graphical horseshoe prior. arXiv preprint. Available at arXiv:2302.06423.

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

Key words and phrases. Bayesian graphical models, cancer genomics, expectation conditional maximisation, gene networks, genomics, graphical horseshoe, horseshoe prior, high-dimensional inference, integrative analysis, multiomics, network models.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751 https://doi.org/10.1093/biomet/asq017

Chen, H. and Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinform.* **5** 1–13.

Conway, J. R., Lex, A. and Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics*.

Danaher, P., Wang, P. and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 373–397. MR3164871 https://doi.org/10.1111/rssb.12033

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

Dobrovolsky, V. N., Bowyer, J. F., Pabarcus, M. K., Heflich, R. H., Williams, L. D., Doerge, D. R., Arvidsson, B., Bergquist, J. and Casida, J. E. (2005). Effect of arylformamidase (kynurenine formamidase) gene inactivation in mice on enzymatic activity, kynurenine pathway metabolites and phenotype. *Biochim. Biophys. Acta, Gen. Subj.* **1724** 163–172.

Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R. et al. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343** 1246949.

Fairfax, B. P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F. O. et al. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44** 502–510. https://doi.org/10.1038/ng.2205

Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.* **3** 521–541. MR2750671 https://doi.org/10.1214/08-AOAS215

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441. https://doi.org/10.1093/biostatistics/kxm045

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1–15. MR2804206 https://doi.org/10.1093/biomet/asq060

Hao, L., Sakurai, A., Watanabe, T., Sorensen, E., Nidom, C. A., Newton, M. A., Ahlquist, P. and Kawaoka, Y. (2008). Drosophila RNAi screen identifies host genes important for influenza virus replication. *Nature* **454** 890–893. https://doi.org/10.1038/nature07151

Hugill, A. J., Stewart, M. E., Yon, M. A., Probert, F., Cox, I. J., Hough, T. A., Scudamore, C. L., Bentley, L., Wall, G. et al. (2015). Loss of arylformamidase with reduced thymidine kinase expression leads to impaired glucose tolerance. *Biol. Open* **4** 1367–1375.

Karczewski, K. J. and Snyder, M. P. (2018). Integrative omics for health and disease. *Nat. Rev. Genet.* **19** 299–310.

Kim, S., Becker, J., Bechheim, M., Kaiser, V., Noursadeghi, M., Fricker, N., Beier, E., Klaschik, S., Boor, P. et al. (2014). Characterizing the genetic basis of innate immune response in TLR4-activated human monocytes. *Nat. Commun.* **5** 1–7.

Kirkpatrick, S., Gelatt, C. D. Jr. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220** 671–680. MR0702485 https://doi.org/10.1126/science.220.4598.671

Kook, J. H., Vaughn, K. A., DeMaster, D. M., Ewing-Cobbs, L. and Vannucci, M. (2021). BVAR-connect: A variational Bayes approach to multi-subject vector autoregressive models for inference on brain connectivity networks. *Neuroinformatics* **19** 39–56.

Kyewski, B. and Klein, L. (2006). A central role for central tolerance. *Annu. Rev. Immunol.* **24** 571–606. https://doi.org/10.1146/annurev.immunol.23.021704.115601

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

Lee, M. N., Ye, C., Villani, A.-C., Raj, T., Li, W., Eisenhaure, T. M., Imboywa, S. H., Chipendo, P. I., Ran, F. A. et al. (2014). Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343** 1246980.

Li, Y., Craig, B. A. and Bhadra, A. (2019). The graphical horseshoe estimator for inverse covariance matrices. *J. Comput. Graph. Statist.* **28** 747–757. MR4007755 https://doi.org/10.1080/10618600.2019.1575744

Li, Z., Mccormick, T. and Clark, S. (2019). Bayesian joint spike-and-slab graphical lasso. In *International Conference on Machine Learning* 3877–3885. PMLR.

Lingjærde, C., Fairfax, B. P., Richardson, S. and Ruffieux, H. (2024a). Supplement D to “Scalable multiple network inference with the joint graphical horseshoe.” https://doi.org/10.1214/23-AOAS1863SUPPD

Lingjærde, C., Fairfax, B. P., Richardson, S. and Ruffieux, H. (2024b). Supplement A to “Scalable multiple network inference with the joint graphical horseshoe.” https://doi.org/10.1214/23-AOAS1863SUPPA

LINGJÆRDE, C., FAIRFAX, B. P., RICHARDSON, S. and RUFFIEUX, H. (2024c). Supplement B to "Scalable multiple network inference with the joint graphical horseshoe." https://doi.org/10.1214/23-AOAS1863SUPPB

LINGJÆRDE, C., FAIRFAX, B. P., RICHARDSON, S. and RUFFIEUX, H. (2024d). Supplement C to "Scalable multiple network inference with the joint graphical horseshoe." https://doi.org/10.1214/23-AOAS1863SUPPC

LINGJÆRDE, C. and RICHARDSON, S. (2023). StabJGL: A stability approach to sparsity and similarity selection in multiple network reconstruction. arXiv preprint. Available at arXiv:2306.03212.

LISTON, A., LESAGE, S., WILSON, J., PELTONEN, L. and GOODNOW, C. C. (2003). Aire regulates negative selection of organ-specific T cells. *Nat. Immunol.* **4** 350–354.

LIU, R., GAO, J., YANG, Y., QIU, R., ZHENG, Y., HUANG, W., ZENG, Y., HOU, Y., WANG, S. et al. (2018). PHD finger protein 1 (PHF1) is a novel reader for histone H4R3 symmetric dimethylation and coordinates with PRMT5–WDR77/CRL4B complex to promote tumorigenesis. *Nucleic Acids Res.* **46** 6608–6626.

MAKALIC, E. and SCHMIDT, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Process. Lett.* **23** 179–182.

MATHIS, D. and BENOIST, C. (2007). A decade of AIRE. *Nat. Rev., Immunol.* **7** 645–650. https://doi.org/10.1038/nri2136

MCLACHLAN, G. J. and BASFORD, K. E. (1988). *Mixture Models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs* **84**. Dekker, New York. MR0926484

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

MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503 https://doi.org/10.1093/biomet/80.2.267

NI, Y., BALADANDAYUTHAPANI, V., VANNUCCI, M. and STINGO, F. C. (2022). Bayesian graphical models for modern biological applications. *Stat. Methods Appl.* **31** 197–225. MR4426829 https://doi.org/10.1007/s10260-021-00572-8

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

PETERSON, P., ORG, T. and REBANE, A. (2008). Transcriptional regulation by AIRE: Molecular mechanisms of central tolerance. *Nat. Rev., Immunol.* **8** 948–957. https://doi.org/10.1038/nri2450

PIIRONEN, J. and VEHTARI, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Stat.* **11** 5018–5051. MR3738204 https://doi.org/10.1214/17-EJS1337SI

POLSON, N. G. and SCOTT, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Stat.* **9** 105.

ROČKOVÁ, V. and GEORGE, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *J. Amer. Statist. Assoc.* **109** 828–846. MR3223753 https://doi.org/10.1080/01621459.2013.869223

ROČKOVÁ, V. and GEORGE, E. I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.* **113** 431–444. MR3803476 https://doi.org/10.1080/01621459.2016.1260469

RUBIN, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9** 130–134. MR0600538

RUFFIEUX, H., DAVISON, A. C., HAGER, J., INSHAW, J., FAIRFAX, B. P., RICHARDSON, S. and BOTTOLO, L. (2020). A global-local approach for detecting hotspots in multiple-response regression. *Ann. Appl. Stat.* **14** 905–928. MR4117834 https://doi.org/10.1214/20-AOAS1332

RUFFIEUX, H., FAIRFAX, B. P., NASSIRI, I., VIGORITO, E., WALLACE, C., RICHARDSON, S. and BOTTOLO, L. (2021). EPISPOT: An epigenome-driven approach for detecting and interpreting hotspots in molecular QTL studies. *Am. J. Hum. Genet.* **108** 983–1000. https://doi.org/10.1016/j.ajhg.2021.04.010

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

SOMEREN, E. V., WESSELS, L. F., BACKER, E. and REINDERS, M. J. (2002). Genetic network modeling. *Pharmacogenomics J.* **3** 507–525.

R CORE TEAM (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

VAN DE WIEL, M. A., TE BEEST, D. E. and MÜNCH, M. M. (2019). Learning from a lot: Empirical Bayes for high-dimensional model-based prediction. *Scand. J. Stat.* **46** 2–25. MR3915265 https://doi.org/10.1111/sjos.12335

VAN DER PAS, S. L., KLEIJN, B. J. K. and VAN DER VAART, A. W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electron. J. Stat.* **8** 2585–2618. MR3285877 https://doi.org/10.1214/14-EJS962

WANG, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Anal.* **7** 867–886. MR3000017 https://doi.org/10.1214/12-BA729

WANG, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Anal.* **10** 351–377. MR3420886 https://doi.org/10.1214/14-BA916

WANG, L., HUANG, Y., WANG, X. and CHEN, Y. (2019). Label-free LC-MS/MS proteomics analyses reveal proteomic changes accompanying MSTN KO in C2C12 cells. *BioMed Res. Int.* **2019**.

YANG, X., GAN, L., NARISETTY, N. N. and LIANG, F. (2021). GemBag: Group estimation of multiple Bayesian graphical models. *J. Mach. Learn. Res.* **22** 54. MR4253747

YAO, C., JOEHANES, R., JOHNSON, A. D., HUAN, T., LIU, C., FREEDMAN, J. E., MUNSON, P. J., HILL, D. E., VIDAL, M. et al. (2017). Dynamic role of trans regulation of gene expression in relation to complex traits. *Am. J. Hum. Genet.* **100** 571–580.

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 1924–1946
https://doi.org/10.1214/23-AOAS1864

JOINT MIXED MEMBERSHIP MODELING OF MULTIVARIATE LONGITUDINAL AND SURVIVAL DATA FOR LEARNING THE INDIVIDUALIZED DISEASE PROGRESSION

BY YUYANG HE[1,a], XINYUAN SONG[1,b] AND KAI KANG[2,c]

[1]*Department of Statistics, The Chinese University of Hong Kong,* [a]*yuyanghe@link.cuhk.edu.hk,* [b]*xysong@cuhk.edu.hk*
[2]*Department of Statistics, Sun Yat-sen University,* [c]*kangk5@mail.sysu.edu.cn*

Patients with Alzheimer's disease (AD) often exhibit substantial heterogeneity in disease progression due to multiple genetic causes for such a complex disease. Investigating diverse subtypes of neurodegeneration and individualized disease progression is essential for early diagnosis and precision medicine. In this article we present a novel joint mixed membership model for multivariate longitudinal AD-related biomarkers and time of AD diagnosis. Unlike conventional finite mixture models that assign each subject a single subgroup membership, the proposed model assigns partial membership across subgroups, allowing subjects to lie between two or more subgroups. This flexible structure enables individualized disease progression and facilitates the identification of clinically meaningful neurological statuses often elusive in current mixed effects models. We employ a spline-based trajectory model to characterize complex and possibly nonlinear patterns of multiple longitudinal clinical markers. A Cox model is then used to examine the effects of time-variant risk factors on the hazard of developing AD. We develop a Bayesian method coupled with efficient Markov chain Monte Carlo sampling schemes to perform statistical inference. The proposed approach is assessed through extensive simulation studies and an application to the Alzheimer's Disease Neuroimaging Initiative study, showing a better performance in AD diagnosis than existing joint models.

REFERENCES

AIROLDI, E. M., FIENBERG, S. E., JOUTARD, C. and LOVE, T. M. (2006). Discovering latent patterns with hierarchical Bayesian mixed-membership models. In *Data Mining Patterns*: *New Methods and Applications*. *Idea Group Inc* 240–275.

ANDRINOPOULOU, E.-R., EILERS, P. H. C., TAKKENBERG, J. J. M. and RIZOPOULOS, D. (2018). Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using P-splines. *Biometrics* **74** 685–693. MR3825355 https://doi.org/10.1111/biom.12814

BALTHAZAR, M. L., YASUDA, C. L., CENDES, F. and DAMASCENO, B. P. (2010). Learning, retrieval, and recognition are compromised in a MCI and mild AD: Are distinct episodic memory processes mediated by the same anatomical structures? *J. Int. Neuropsychol. Soc.* **16** 205–209.

BARRETT, J., DIGGLE, P., HENDERSON, R. and TAYLOR-ROBINSON, D. (2015). Joint modelling of repeated measurements and time-to-event outcomes: Flexible model specification and exact likelihood inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 131–148. MR3299402 https://doi.org/10.1111/rssb.12060

BROWN, E. R. and IBRAHIM, J. G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59** 221–228. MR1987388 https://doi.org/10.1111/1541-0420.00028

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

CONGDON, P. (2014). *Applied Bayesian Modelling*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3221807 https://doi.org/10.1002/9781118895047

CORDER, E. H., SAUNDERS, A. M., STRITTMATTER, W. J. et al. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261** 921–923.

Key words and phrases. Mixed membership model, longitudinal data, MCMC methods, survival data.

COSENTINO, S., SCARMEAS, N., HELZNER, E. et al. (2008). APOE-ϵ4 allele predicts faster cognitive decline in mild Alzheimer disease. *Neurology* **70** 1842–1849.

EROSHEVA, E. A. (2002). *Grade of Membership and Latent Structure Models with Application to Disability Survey Data*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Carnegie Mellon University. MR2703653

EROSHEVA, E. A., FIENBERG, S. E. and JOUTARD, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *Ann. Appl. Stat.* **1** 502–537. MR2415745 https://doi.org/10.1214/07-AOAS126

ESTÉVEZ-GONZÁLEZ, A., KULISEVSKY, J., BOLTES, A., OTERMÍN, P. and GARCÍA-SÁNCHEZ, C. (2003). Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer's disease: Comparison with mild cognitive impairment and normal aging. *Int. J. Geriatr. Psychiatry* **18** 1021–1028.

FENG, X.-N., WANG, G.-C., WANG, Y.-F. and SONG, X.-Y. (2015). Structure detection of semiparametric structural equation models with Bayesian adaptive group lasso. *Stat. Med.* **34** 1527–1547. MR3334674 https://doi.org/10.1002/sim.6410

FITZPATRICK, A. L., KULLER, L. H., IVES, D. G., LOPEZ, O. L., JAGUST, W., BREITNER, J. C. S., JONES, B., LYKETSOS, C. and DULBERG, C. (2004). Incidence and prevalence of dementia in the cardiovascular health study. *J. Amer. Geriatr. Soc.* **52** 195–204. https://doi.org/10.1111/j.1532-5415.2004.52058.x

FRÜHWIRTH-SCHNATTER, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J. Amer. Statist. Assoc.* **96** 194–209. MR1952732 https://doi.org/10.1198/016214501750333063

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. *Springer Series in Statistics*. Springer, New York. MR2265601

GORMLEY, I. C. and MURPHY, T. B. (2009). A grade of membership model for rank data. *Bayesian Anal.* **4** 265–295. MR2507364 https://doi.org/10.1214/09-BA410

GU, Y., EROSHEVA, E. A., XU, G. and DUNSON, D. B. (2023). Dimension-grouped mixed membership models for multivariate categorical data. *J. Mach. Learn. Res.* **24** 88. MR4582510

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109. MR3363437 https://doi.org/10.1093/biomet/57.1.97

HE, Y., SONG, X. and KANG, K. (2024). Supplement to "Joint mixed membership modeling of multivariate longitudinal and survival data for learning the individualized disease progression." https://doi.org/10.1214/23-AOAS1864SUPP

HEAGERTY, P. J. and ZHENG, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61** 92–105. MR2135849 https://doi.org/10.1111/j.0006-341X.2005.030814.x

HOLMES, C. C. and MALLICK, B. K. (2003). Generalized nonlinear modeling with multivariate free-knot regression splines. *J. Amer. Statist. Assoc.* **98** 352–368. MR1995711 https://doi.org/10.1198/016214503000143

IBRAHIM, J. G., CHEN, M.-H. and SINHA, D. (2001). *Bayesian Survival Analysis*. *Springer Series in Statistics*. Springer, New York. MR1876598 https://doi.org/10.1007/978-1-4757-3447-8

IBRAHIM, J. G., CHEN, M.-H. and SINHA, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statist. Sinica* **14** 863–883. MR2087976

JACK, C. R., KNOPMAN, D. S., JAGUST, W. J. et al. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* **9** 119–128.

KANG, K., CAI, J., SONG, X. and ZHU, H. (2019). Bayesian hidden Markov models for delineating the pathology of Alzheimer's disease. *Stat. Methods Med. Res.* **28** 2112–2124. MR3977095 https://doi.org/10.1177/0962280217748675

KANG, K. and SONG, X. (2022). Consistent estimation of a joint model for multivariate longitudinal and survival data with latent variables. *J. Multivariate Anal.* **187** 104827. MR4319408 https://doi.org/10.1016/j.jmva.2021.104827

KANG, K. and SONG, X. Y. (2023). Joint modeling of longitudinal imaging and survival data. *J. Comput. Graph. Statist.* **32** 402–412. MR4592919 https://doi.org/10.1080/10618600.2022.2102027

KANTARCI, K., GUNTER, J. L., TOSAKULWONG, N. et al. (2013). Focal hemosiderin deposits and β-amyloid load in the ADNI cohort. *Alzheimer's Dement.* **9** S116–S123.

KLEIN ENTINK, R. H., FOX, J.-P. and VAN DEN HOUT, A. (2011). A mixture model for the joint analysis of latent developmental trajectories and survival. *Stat. Med.* **30** 2310–2325. MR2830010 https://doi.org/10.1002/sim.4266

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

LEE, S. and GU, Y. (2022). New Paradigm of identifiable general-response cognitive diagnostic models: Beyond categorical data. Available at: https://stat.columbia.edu/~yuqigu/assets/pdf/GR_CDM_1018.pdf.

LI, C., XIAO, L. and LUO, S. (2022). Joint model for survival and multivariate sparse functional data with application to a study of Alzheimer's Disease. *Biometrics* **78** 435–447. MR4450566 https://doi.org/10.1111/biom.13427

LI, K. and LUO, S. (2017). Functional joint model for longitudinal and time-to-event data: An application to Alzheimer's disease. *Stat. Med.* **36** 3560–3572. MR3696509 https://doi.org/10.1002/sim.7381

MAIER, M. (2014). DirichletReg: Dirichlet regression for compositional data in R.

MANRIQUE-VALLIER, D. (2014). Longitudinal mixed membership trajectory models for disability survey data. *Ann. Appl. Stat.* **8** 2268–2291. MR3292497 https://doi.org/10.1214/14-AOAS769

MANRIQUE-VALLIER, D. and FIENBERG, S. E. (2008). Population size estimation using individual level mixture models. *Biom. J.* **50** 1051–1063. MR2649394 https://doi.org/10.1002/bimj.200810448

MCLACHLAN, G. J. and BASFORD, K. E. (1988). *Mixture Models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs* **84**. Dekker, New York. MR0926484

MEHTA, K. M. and YEO, G. W. (2017). Systematic review of dementia prevalence and incidence in United States race/ethnic populations. *Alzheimer's Dement.* **13** 72–83.

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

PARK, L. Q., GROSS, A. L., MCLAREN, D. G., PA, J., JOHNSON, J. K. and MITCHELL, M. (2012). Confirmatory factor analysis of the ADNI neuropsychological battery. *Brain Imaging Behav.* **6** 528–539.

PETERSEN, R. C. (2004). Mild cognitive impairment as a diagnostic entity. *J. Intern. Med.* **256** 183–194. https://doi.org/10.1111/j.1365-2796.2004.01388.x

RAFTERY, A. E., NEWTON, M. A., SATAGOPAN, J. M. and KRIVITSKY, P. N. (2006). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Stat.* **8** 1–45.

RIZOPOULOS, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67** 819–829. MR2829256 https://doi.org/10.1111/j.1541-0420.2010.01546.x

RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.* **11** 735–757. MR1944261 https://doi.org/10.1198/106186002321018768

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

SONG, F., CHU, J., MA, S. and WEI, Y. (2023). Survival mixed membership blockmodel. *J. Amer. Statist. Assoc.* 1–19.

SONG, X.-Y. and LU, Z.-H. (2010). Semiparametric latent variable models with Bayesian P-splines. *J. Comput. Graph. Statist.* **19** 590–608. MR2732494 https://doi.org/10.1198/jcgs.2010.09094

SUN, J., HERAZO-MAYA, J. D., MOLYNEAUX, P. L., MAHER, T. M., KAMINSKI, N. and ZHAO, H. (2019). Regularized latent class model for joint analysis of high-dimensional longitudinal biomarkers and a time-to-event outcome. *Biometrics* **75** 69–77. MR3953708 https://doi.org/10.1111/biom.12964

TSENG, Y.-K., HSIEH, F. and WANG, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92** 587–603. MR2202648 https://doi.org/10.1093/biomet/92.3.587

VIRTANEN, S. and GIROLAMI, M. (2015). Ordinal mixed membership models. In *International Conference on Machine Learning* 588–596.

WANG, J., HOEKSTRA, J. G., ZUO, C., COOK, T. J. and ZHANG, J. (2013). Biomarkers of Parkinson's disease: Current status and future perspectives. *Drug Discov. Today* **18** 155–162.

WANG, Q. and WANG, Y. (2024). Multilayer exponential family factor models for integrative analysis and learning disease progression. *Biostatistics* **25** 203–219. MR4678542 https://doi.org/10.1093/biostatistics/kxac042

WANG, Y. (2019). Convergence rates of latent topic models under relaxed identifiability conditions. *Electron. J. Stat.* **13** 37–66. MR3896145 https://doi.org/10.1214/18-ejs1516

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

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

ZHENG, G. (2021). A mixed membership Rasch model Ph.D. thesis, Carnegie Mellon Univ., Pittsburgh, PA.

ZOU, H., ZENG, D., XIAO, L. and LUO, S. (2023). Bayesian inference and dynamic prediction for multivariate longitudinal and survival data. *Ann. Appl. Stat.* **17** 2574–2595. MR4637681 https://doi.org/10.1214/23-aoas1733

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 1947–1964
https://doi.org/10.1214/23-AOAS1865
© Institute of Mathematical Statistics, 2024

OUTCOME-GUIDED DISEASE SUBTYPING BY GENERATIVE MODEL AND WEIGHTED JOINT LIKELIHOOD IN TRANSCRIPTOMIC APPLICATIONS

BY YUJIA LI[a], PENG LIU[b], WENJIA WANG[c], WEI ZONG[d], YUSI FANG[e], ZHAO REN[f], LU TANG[g], JUAN C. CELEDÓN[h], STEFFI OESTERREICH[i] AND GEORGE C. TSENG[j]

University of Pittsburgh, [a]*yul178@pitt.edu,* [b]*pel67@pitt.edu,* [c]*wew89@pitt.edu,* [d]*wez97@pitt.edu,* [e]*yuf31@pitt.edu,* [f]*zren@pitt.edu,* [g]*lutang@pitt.edu,* [h]*celedonj@pitt.edu,* [i]*oesterreichs@upmc.edu,* [j]*ctseng@pitt.edu*

With advances in high-throughput technology, molecular disease subtyping by high-dimensional omics data has been recognized as an effective approach for identifying subtypes of complex diseases with distinct disease mechanisms and prognoses. Conventional cluster analysis takes omics data as input and generates patient clusters with similar gene expression pattern. The omics data, however, usually contain multifaceted cluster structures that can be defined by different sets of genes. If the gene set associated with irrelevant clinical variables (e.g., sex or age) dominates the clustering process, the resulting clusters may not capture clinically meaningful disease subtypes. This motivates the development of a clustering framework with guidance from a prespecified disease outcome, such as lung function measurement or survival, in this paper. We propose two disease subtyping methods by omics data with outcome guidance using a generative model or a weighted joint likelihood. Both methods connect an outcome association model and a disease subtyping model by a latent variable of cluster labels. Compared to the generative model, weighted joint likelihood contains a data-driven weight parameter to balance the likelihood contributions from outcome association and gene cluster separation, which improves generalizability in independent validation but requires heavier computing. Extensive simulations and two real applications in lung disease and triple-negative breast cancer demonstrate superior disease subtyping performance of the outcome-guided clustering methods in terms of disease subtyping accuracy, gene selection and outcome association. Unlike existing clustering methods, the outcome-guided disease subtyping framework creates a new precision medicine paradigm to directly identify patient subgroups with clinical association.

REFERENCES

BAIR, E. and TIBSHIRANI, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.* **2** E108. https://doi.org/10.1371/journal.pbio.0020108

CHANG, W., WAN, C., ZANG, Y., ZHANG, C. and CAO, S. (2020). Supervised clustering of high-dimensional data using regularized mixture modeling. *Brief. Bioinform.* **22** bbaa291.

CHUNG, K. F. (2001). Cytokines in chronic obstructive pulmonary disease. *Eur. Respir. J.* **18** 50s–59s.

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

DEAN, N. and RAFTERY, A. E. (2010). Latent class analysis variable selection. *Ann. Inst. Statist. Math.* **62** 11–35. MR2577437 https://doi.org/10.1007/s10463-009-0258-9

DESANTIS, S. M., HOUSEMAN, E. A., COULL, B. A., NUTT, C. L. and BETENSKY, R. A. (2012). Supervised Bayesian latent class models for high-dimensional data. *Stat. Med.* **31** 1342–1360. MR2925053 https://doi.org/10.1002/sim.4448

DESANTIS, S. M., HOUSEMAN, E. A., COULL, B. A., STEMMER-RACHAMIMOV, A. and BETENSKY, R. A. (2008). A penalized latent class model for ordinal data. *Biostatistics* **9** 249–262. https://doi.org/10.1093/biostatistics/kxm026

FOP, M. and MURPHY, T. B. (2018). Variable selection methods for model-based clustering. *Stat. Surv.* **12** 18–65. MR3794323 https://doi.org/10.1214/18-SS119

Key words and phrases. Disease subtyping, omics data, high-dimensional cluster analysis, generative model, weighted joint likelihood.

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

FURGAL, A. K. C., SEN, A. and TAYLOR, J. M. G. (2019). Review and comparison of computational approaches for joint longitudinal and time-to-event models. *Int. Stat. Rev.* **87** 393–418. MR3994765 https://doi.org/10.1111/insr.12322

GORMLEY, I. C. and FRÜHWIRTH-SCHNATTER, S. (2019). Mixture of experts models. In *Handbook of Mixture Analysis*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 271–307. CRC Press, Boca Raton, FL. MR3889697

GUO, J., WALL, M. and AMEMIYA, Y. (2006). Latent class regression on latent factors. *Biostatistics* **7** 145–163. https://doi.org/10.1093/biostatistics/kxi046

HOUSEMAN, E. A., COULL, B. A. and BETENSKY, R. A. (2006). Feature-specific penalized latent class analysis for genomic data. *Biometrics* **62** 1062–1070. MR2297677 https://doi.org/10.1111/j.1541-0420.2006.00566.x

HUBERT, L. J. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.

JEMAL, A., SIEGEL, R., WARD, E., HAO, Y., XU, J. and THUN, M. J. (2009). Cancer statistics, 2009. *CA Cancer J. Clin.* **59** 225–249.

LANZA, S. T. and RHOADES, B. L. (2013). Latent class analysis: An alternative perspective on subgroup analysis in prevention and treatment. *Prev. Sci.* **14** 157–168. https://doi.org/10.1007/s11121-011-0201-1

LI, Y., LIU, P., WANG, W., ZONG, W., FANG, Y., REN, Z., TANG, L., CELEDÓN, J. C, OESTERREICH, S. and TSENG, G. C (2024). Supplement to “Outcome-guided disease subtyping by generative model and weighted joint likelihood in transcriptomic applications.” https://doi.org/10.1214/23-AOAS1865SUPP

LI, Y., RAHMAN, T., MA, T., TANG, L. and TSENG, G. C. (2023). A sparse negative binomial mixture model for clustering RNA-seq count data. *Biostatistics* **24** 68–84. MR4522704 https://doi.org/10.1093/biostatistics/kxab025

LI, Y., ZENG, X., LIN, C.-W. and TSENG, G. C. (2022). Simultaneous estimation of cluster number and feature sparsity in high-dimensional cluster analysis. *Biometrics* **78** 574–585. MR4450577 https://doi.org/10.1111/biom.13449

LIN, B., BAI, L., WANG, S. and LIN, H. (2021). The association of systemic interleukin 6 and interleukin 10 levels with sarcopenia in elderly patients with chronic obstructive pulmonary disease. *Int. J. Gen. Med.* **14** 5893–5902.

LIN, H., TURNBULL, B. W., MCCULLOCH, C. E. and SLATE, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *J. Amer. Statist. Assoc.* **97** 53–65. MR1947272 https://doi.org/10.1198/016214502753479220

LOCK, E. F. and DUNSON, D. B. (2013). Bayesian consensus clustering. *Bioinformatics* **29** 2610–2616. https://doi.org/10.1093/bioinformatics/btt425

OGAWA, Y., DURU, E. A. and AMEREDES, B. T. (2008). Role of IL-10 in the resolution of airway inflammation. *Curr. Mol. Med.* **8** 437–445. https://doi.org/10.2174/156652408785160907

OSHI, M., LE, L., ANGARITA, F. A., TOKUMARU, Y., YAN, L., MATSUYAMA, R., ENDO, I. and TAKABE, K. (2021). Association of allograft rejection response score with biological cancer aggressiveness and with better survival in triple-negative breast cancer (TNBC).

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

PENCINA, M. J. and D’AGOSTINO, R. B. (2004). Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. *Stat. Med.* **23** 2109–2123.

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

PLANES-LAINE, G., ROCHIGNEUX, P., BERTUCCI, F., CHRÉTIEN, A.-S., VIENS, P., SABATIER, R. and GONÇALVES, A. (2019). PD-1/PD-L1 targeting in breast cancer: The first clinical evidences are emerging—a literature review. *Cancers* **11** 1033.

PROUST-LIMA, C., SÉNE, M., TAYLOR, J. M. G. and JACQMIN-GADDA, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Stat. Methods Med. Res.* **23** 74–90. MR3190688 https://doi.org/10.1177/0962280212445839

PROUST-LIMA, C. and TAYLOR, J. M. G. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: A joint modeling approach. *Biostatistics* **10** 535–549. https://doi.org/10.1093/biostatistics/kxp009

SCHRÖDER, M. S., CULHANE, A. C., QUACKENBUSH, J. and HAIBE-KAINS, B. (2011). Survcomp: An R/bioconductor package for performance assessment and comparison of survival models. *Bioinformatics* **27** 3206–3208.

SUN, J., HERAZO-MAYA, J. D., MOLYNEAUX, P. L., MAHER, T. M., KAMINSKI, N. and ZHAO, H. (2019). Regularized latent class model for joint analysis of high-dimensional longitudinal biomarkers and a time-to-event outcome. *Biometrics* **75** 69–77. MR3953708 https://doi.org/10.1111/biom.12964

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

TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 411–423. MR1841503 https://doi.org/10.1111/1467-9868.00293

WANG, D.-R., WU, X.-L. and SUN, Y.-L. (2022). Therapeutic targets and biomarkers of tumor immunotherapy: Response versus non-response. *Signal Transduct. Targeted Ther.* **7**.

WITTEN, D. M. and TIBSHIRANI, R. (2010). A framework for feature selection in clustering. *J. Amer. Statist. Assoc.* **105** 713–726. MR2724855 https://doi.org/10.1198/jasa.2010.tm09415

WITTEN, D. M. and TIBSHIRANI, R. (2011). Penalized classification using Fisher's linear discriminant. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 753–772. MR2867457 https://doi.org/10.1111/j.1467-9868.2011.00783.x

ZHAO, X., LIU, J., GE, S., CHEN, C., LI, S., WU, X., FENG, X., WANG, Y. and CAI, D. (2019). Saikosaponin A inhibits breast cancer by regulating Th1/Th2 balance. *Frontiers in Pharmacology* **10** 624.

ZHOU, H., PAN, W. and SHEN, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electron. J. Stat.* **3** 1473–1496. MR2578834 https://doi.org/10.1214/09-EJS487

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

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 1965–1987
https://doi.org/10.1214/23-AOAS1866

LEARNING AND FORECASTING OF AGE-SPECIFIC PERIOD MORTALITY VIA B-SPLINE PROCESSES WITH LOCALLY-ADAPTIVE DYNAMIC COEFFICIENTS

BY FEDERICO PAVONE[1,a], SIRIO LEGRAMANTI[2,c] AND DANIELE DURANTE[1,b]

[1] *Department of Decision Sciences, Bocconi University,* [a] *federico.pavone@phd.unibocconi.it,* [b] *daniele.durante@unibocconi.it*
[2] *Department of Economics, University of Bergamo,* [c] *sirio.legramanti@unibg.it*

Although the analysis of human mortality has a well-established history, the attempt to accurately forecast future death-rate patterns for different age groups and time horizons still attracts active research. Such a predictive focus has motivated an increasing shift toward more flexible representations of age-specific period mortality trajectories at the cost of reduced interpretability. Although this perspective has led to successful predictive strategies, the inclusion of interpretable structures in modeling of human mortality can be, in fact, beneficial for improving forecasts. We pursue this direction via a novel B-spline process with locally-adaptive dynamic coefficients. Such a process outperforms state-of-the-art forecasting strategies by explicitly incorporating the core structures of period mortality within an interpretable formulation which enables inference on age-specific mortality trends and the corresponding rates of change across time. This is obtained by modeling the age-specific death counts via a Poisson log-normal model parameterized through a linear combination of B-spline bases with dynamic coefficients that characterize time changes in mortality rates via suitably defined stochastic differential equations. While flexible, the resulting formulation can be accurately approximated by a Gaussian state-space model that facilitates closed-form Kalman filtering, smoothing and forecasting, for both the trends of the spline coefficients and the corresponding first derivatives, which measure rates of change in mortality for different age groups. As illustrated in applications to mortality data from different countries, the proposed model outperforms state-of-the-art methods, both in point forecasts and in calibration of predictive intervals. Moreover, it unveils substantial differences in mortality patterns across countries and ages, both in the past decades and during the COVID-19 pandemic.

REFERENCES

ALEXOPOULOS, A., DELLAPORTAS, P. and FORSTER, J. J. (2019). Bayesian forecasting of mortality rates by using latent Gaussian models. *J. Roy. Statist. Soc. Ser. A* **182** 689–711. MR3902678 https://doi.org/10.1111/rssa.12422

ALIVERTI, E., MAZZUCO, S. and SCARPA, B. (2022). Dynamic modelling of mortality via mixtures of skewed distribution functions. *J. Roy. Statist. Soc. Ser. A* **185** 1030–1048. MR4463267 https://doi.org/10.1111/rssa.12808

BOOTH, H. and TICKLE, L. (2008). Mortality modelling and forecasting: A review of methods. *Ann. Actuar. Sci.* **3** 3–43.

BROUHNS, N., DENUIT, M. and VERMUNT, J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance Math. Econom.* **31** 373–393. MR1945540 https://doi.org/10.1016/S0167-6687(02)00185-3

CAIRNS, A. J., BLAKE, D. and DOWD, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *J. Risk Insur.* **73** 687–718.

CAIRNS, A. J. G., BLAKE, D., DOWD, K., COUGHLAN, G. D., EPSTEIN, D., ONG, A. and BALEVICH, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *N. Am. Actuar. J.* **13** 1–35. MR2496489 https://doi.org/10.1080/10920277.2009.10597538

CAMARDA, C. G. (2019). Smooth constrained mortality forecasting. *Demogr. Res.* **41** 1091–1130.

Key words and phrases. B-spline, death rate, Kalman filter, mortality forecasting, nested Gaussian process.

CASE, A. and DEATON, A. (2021). *Deaths of Despair and the Future of Capitalism*. Princeton Univ. Press, Princeton.

CHOPIN, N. and PAPASPILIOPOULOS, O. (2020). *An Introduction to Sequential Monte Carlo*. *Springer Series in Statistics*. Springer, Cham. MR4215639 https://doi.org/10.1007/978-3-030-47845-2

CONTI, S., FARCHI, G. and PRATI, S. (1994). AIDS as a leading cause of death among young adults in Italy. *Eur. J. Epidemiol.* **10** 669–673. https://doi.org/10.1007/BF01719279

CONTI, S., MASOCCO, M., FARCHI, G., REZZA, G. and TOCCACELI, V. (1997). Premature mortality in Italy during the first decade of the AIDS epidemic: 1984–1993. *Int. J. Epidemiol.* **26** 873–879. https://doi.org/10.1093/ije/26.4.873

CURRIE, I. D. (2016). On fitting generalized linear and non-linear models of mortality. *Scand. Actuar. J.* 4 356–383. MR3435188 https://doi.org/10.1080/03461238.2014.928230

CURRIE, I. D., DURBAN, M. and EILERS, P. H. C. (2004). Smoothing and forecasting mortality rates. *Stat. Model.* **4** 279–298. MR2086492 https://doi.org/10.1191/1471082X04st080oa

CZADO, C., DELWARDE, A. and DENUIT, M. (2005). Bayesian Poisson log-bilinear mortality projections. *Insurance Math. Econom.* **36** 260–284. MR2152844 https://doi.org/10.1016/j.insmatheco.2005.01.001

DELLAPORTAS, P., SMITH, A. F. M. and STAVROPOULOS, P. (2001). Bayesian analysis of mortality data. *J. Roy. Statist. Soc. Ser. A* **164** 275–291. MR1830699 https://doi.org/10.1111/1467-985X.00202

DELWARDE, A., DENUIT, M. and EILERS, P. (2007). Smoothing the Lee–Carter and Poisson log-bilinear models for mortality forecasting: A penalized log-likelihood approach. *Stat. Model.* **7** 29–48. MR2749822 https://doi.org/10.1177/1471082X0600700103

DREFAHL, S., AHLBOM, A. and MODIG, K. (2014). Losing ground-Swedish life expectancy in a comparative perspective. *PLoS ONE* **9** e88357.

DURBIN, J. and KOOPMAN, S. J. (2012). *Time Series Analysis by State Space Methods*, 2nd ed. *Oxford Statistical Science Series* **38**. Oxford Univ. Press, Oxford. MR3014996 https://doi.org/10.1093/acprof:oso/9780199641178.001.0001

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

GINSBORG, P. (1990). *A History of Contemporary Italy*: 1943-80. Penguin, UK.

GLEI, D. A. (2022). The US midlife mortality crisis continues: Excess cause-specific mortality during 2020. *Amer. J. Epidemiol.* **191** 1677–1686. https://doi.org/10.1093/aje/kwac055

GOLDSTEIN, J. R. and LEE, R. D. (2020). Demographic perspectives on the mortality of COVID-19 and other epidemics. *Proc. Natl. Acad. Sci. USA* **117** 22035–22041.

HABERMAN, S. and RENSHAW, A. (2011). A comparative study of parametric mortality projection models. *Insurance Math. Econom.* **48** 35–55. MR2796688 https://doi.org/10.1016/j.insmatheco.2010.09.003

HELIGMAN, L. and POLLARD, J. H. (1980). The age pattern of mortality. *J. Inst. Actuar.* **107** 49–80.

HELSKE, J. (2017). KFAS: Exponential family state space models in R. *J. Stat. Softw.* **78** 1–39. https://doi.org/10.18637/jss.v078.i10

HO, J. Y. and PRESTON, S. H. (2010). US mortality in an international context: Age variations. *Popul. Dev. Rev.* **36** 749–773.

HOLFORD, T. R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics* **39** 311–324. MR0714415 https://doi.org/10.2307/2531004

HUNT, A. and BLAKE, D. (2021). On the structure and classification of mortality models. *N. Am. Actuar. J.* **25** S215–S234. MR4223257 https://doi.org/10.1080/10920277.2019.1649156

HYNDMAN, R. J., BOOTH, H., TICKLE, L. and MAINDONALD, J. (2014). Demography: Forecasting mortality, fertility, migration and population data. R package version 1.18. https://CRAN.R-project.org/package=demography.

HYNDMAN, R. J., BOOTH, H. and YASMEEN, F. (2013). Coherent mortality forecasting: The product-ratio method with functional time series models. *Demography* **50** 261–283. https://doi.org/10.1007/s13524-012-0145-5

HYNDMAN, R. J. and SHAHID ULLAH, MD. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Comput. Statist. Data Anal.* **51** 4942–4956. MR2364551 https://doi.org/10.1016/j.csda.2006.07.028

JUUL, F. E., JODAL, H. C., BARUA, I., REFSUM, E., OLSVIK, Ø., HELSINGEN, L. M., LØBERG, M., BRETTHAUER, M., KALAGER, M. et al. (2022). Mortality in Norway and Sweden during the COVID-19 pandemic. *Scand. J. Public Health* **50** 38–45.

KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82** 35–45. MR3931993

KATZMARZYK, P. T., SALBAUM, J. M. and HEYMSFIELD, S. B. (2020). Obesity, noncommunicable diseases, and COVID-19: A perfect storm. *Am. J. Human Biol.* **32** e23484. https://doi.org/10.1002/ajhb.23484

KJÆGAARD, S., ERGEMEN, Y. E., KALLESTRUP-LAMB, M., OEPPEN, J. and LINDAHL-JACOBSEN, R. (2019). Forecasting causes of death by using compositional data analysis: The case of cancer deaths. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 1351–1370. MR4022816 https://doi.org/10.1111/rssc.12357

KOOPMAN, S. J. and DURBIN, J. (2000). Fast filtering and smoothing for multivariate state space models. *J. Time Series Anal.* **21** 281–296. MR1766960 https://doi.org/10.1111/1467-9892.00186

LAND, K. C. (1986). Methods for national population forecasts: A review. *J. Amer. Statist. Assoc.* **81** 888–901.

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

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

LEE, R. MILLER, T. (2001). Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography* **38** 537–549.

LEE, R. D. and CARTER, L. R. (1992). Modeling and forecasting US mortality. *J. Amer. Statist. Assoc.* **87** 659–671.

LI, N. and LEE, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee–Carter method. *Demography* **42** 575–594.

LI, N., LEE, R. and GERLAND, P. (2013). Extending the Lee–Carter method to model the rotation of age patterns of mortality decline for long-term projections. *Demography* **50** 2037–2051. https://doi.org/10.1007/s13524-013-0232-2

MAZZUCO, S., SCARPA, B. and ZANOTTO, L. (2018). A mortality model based on a mixture distribution function. *Popul. Stud.* **72** 191–200. https://doi.org/10.1080/00324728.2018.1439519

O'HARE, C. and LI, Y. (2012). Explaining young mortality. *Insurance Math. Econom.* **50** 12–25. MR2879021 https://doi.org/10.1016/j.insmatheco.2011.09.005

OSMOND, C. (1985). Using age, period and cohort models to estimate future mortality rates. *Int. J. Epidemiol.* **14** 124–129. https://doi.org/10.1093/ije/14.1.124

PAVONE, F., LEGRAMANTI, S. and DURANTE, D. (2024). Supplement to "Learning and forecasting of age-specific period mortality via B-spline processes with locally-adaptive dynamic coefficients." https://doi.org/10.1214/23-AOAS1866SUPPA, https://doi.org/10.1214/23-AOAS1866SUPPB

PLAT, R. (2009). On stochastic mortality modeling. *Insurance Math. Econom.* **45** 393–404. MR2591315 https://doi.org/10.1016/j.insmatheco.2009.08.006

PRESTON, S. H. and VIERBOOM, Y. C. (2021). Excess mortality in the United States in the 21st century. *Proc. Natl. Acad. Sci.* **118** e2024850118.

RAFTERY, A. E., CHUNN, J. L., GERLAND, P. and SEVČÍKOVÁ, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography* **50** 777–801. https://doi.org/10.1007/s13524-012-0193-x

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

REMUND, A., CAMARDA, C. G. and RIFFE, T. (2018). A cause-of-death decomposition of young adult excess mortality. *Demography* **55** 957–978. https://doi.org/10.1007/s13524-018-0680-9

RENSHAW, A. E. and HABERMAN, S. (2003). Lee–Carter mortality forecasting with age-specific enhancement. *Insurance Math. Econom.* **33** 255–272. MR2039286 https://doi.org/10.1016/S0167-6687(03)00138-0

RENSHAW, A. E. and HABERMAN, S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance Math. Econom.* **38** 556–570.

VAUPEL, J. and LUNDSTROM, H. (1994). Longer life expectancy? Evidence from Sweden of reductions in mortality rates at advanced ages. In *Studies in the Economics of Aging* 79–102. Univ. of Chicago Press, Chicago.

VILLEGAS, A. M., KAISHEV, V. K. and MILLOSSOVICH, P. (2018). StMoMo: An R package for stochastic mortality modelling. *J. Stat. Softw.* **84** 1–38.

WANG, H., PAULSON, K. R., PEASE, S. A., WATSON, S., COMFORT, H., ZHENG, P., ARAVKIN, A. Y., BISIGNANO, C., BARBER, R. M. et al. (2022). Estimating excess mortality due to the COVID-19 pandemic: A systematic analysis of COVID-19-related mortality, 2020–21. *Lancet* **399** 1513–1536.

WANG, P., PANTELOUS, A. A. and VAHID, F. (2023). Multi-population mortality projection: The augmented common factor model with structural breaks. *Int. J. Forecast.* **39** 450–469.

WEN, J., CAIRNS, A. J. and KLEINOW, T. (2021). Fitting multi-population mortality models to socio-economic groups. *Ann. Actuar. Sci.* **15** 144–172.

WIEMERS, E. E., ABRAHAMS, S., ALFAKHRI, M., HOTZ, V. J., SCHOENI, R. F. and SELTZER, J. A. (2020). Disparities in vulnerability to complications from COVID-19 arising from disparities in preexisting conditions in the United States. *Res. Soc. Stratif. Mobil.* **69** 100553. https://doi.org/10.1016/j.rssm.2020.100553

WONG, J. S. T., FORSTER, J. J. and SMITH, P. W. F. (2018). Bayesian mortality forecasting with overdispersion. *Insurance Math. Econom.* **83** 206–221. MR3886500 https://doi.org/10.1016/j.insmatheco.2017.09.023

WOOLF, S. H. and SCHOOMAKER, H. (2019). Life expectancy and mortality rates in the United States, 1959–2017. *JAMA* **322** 1996–2016. https://doi.org/10.1001/jama.2019.16932

ZHU, B. and DUNSON, D. B. (2013). Locally adaptive Bayes nonparametric regression via nested Gaussian processes. *J. Amer. Statist. Assoc.* **108** 1445–1456. MR3174720 https://doi.org/10.1080/01621459.2013.838568

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 1988–2011
https://doi.org/10.1214/23-AOAS1867
© Institute of Mathematical Statistics, 2024

LATENT CONJUNCTIVE BAYESIAN NETWORK: UNIFY ATTRIBUTE HIERARCHY AND BAYESIAN NETWORK FOR COGNITIVE DIAGNOSIS

BY SEUNGHYUN LEE[a] AND YUQI GU[b]

Department of Statistics, Columbia University, [a]*sl4963@columbia.edu,* [b]*yuqi.gu@columbia.edu*

Cognitive diagnostic assessment aims to measure specific knowledge structures in students. To model data arising from such assessments, cognitive diagnostic models with discrete latent variables have gained popularity in educational and behavioral sciences. In a learning context, the latent variables often denote sequentially acquired skill attributes, which is often modeled by the so-called attribute hierarchy method. One drawback of the traditional attribute hierarchy method is that its parameter complexity varies substantially with the hierarchy's graph structure, lacking statistical parsimony. Additionally, arrows among the attributes do not carry an interpretation of statistical dependence. Motivated by these, we propose a new family of *latent conjunctive Bayesian networks* (LCBNs), which rigorously unify the attribute hierarchy method for sequential skill mastery and the Bayesian network model in statistical machine learning. In an LCBN the latent graph not only retains the hard constraints on skill prerequisites as an attribute hierarchy but also encodes nice conditional independence interpretation as a Bayesian network. LCBNs are identifiable, interpretable, and parsimonious statistical tools to diagnose students' cognitive abilities from assessment data. We propose an efficient two-step EM algorithm for structure learning and parameter estimation in LCBNs and establish the consistency of this procedure. Application of our method to an international educational assessment dataset gives interpretable findings of cognitive diagnosis.

REFERENCES

BALAMUTA, J. J. and CULPEPPER, S. A. (2022). Exploratory restricted latent class models with monotonicity requirements under Pòlya-gamma data augmentation. *Psychometrika* **87** 903–945. MR4476253 https://doi.org/10.1007/s11336-021-09815-9

BEERENWINKEL, N., ERIKSSON, N. and STURMFELS, B. (2006). Evolution on distributive lattices. *J. Theoret. Biol.* **242** 409–420. MR2272562 https://doi.org/10.1016/j.jtbi.2006.03.013

BEERENWINKEL, N., ERIKSSON, N. and STURMFELS, B. (2007). Conjunctive Bayesian networks. *Bernoulli* **13** 893–909. MR2364218 https://doi.org/10.3150/07-BEJ6133

BEERENWINKEL, N., RAHNENFÜHRER, J., DÄUMER, M., HOFFMANN, D., KAISER, R., SELBIG, J. and LENGAUER, T. (2005). Learning multiple evolutionary pathways from cross-sectional data. *J. Comput. Biol.* **12** 584–598.

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

BRIGGS, D. C. and ALONZO, A. C. (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. In *Learning Progressions in Science* 293–316. Brill, Leiden.

CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189 https://doi.org/10.1093/biomet/asn034

CHEN, Y. LI, X., LIU, J. and YING, Z. (2018). Recommendation system for adaptive learning. *Appl. Psychol. Meas.* **42** 24–41.

DE LA TORRE, J. (2011). The generalized DINA model framework. *Psychometrika* **76** 179–199. MR2788881 https://doi.org/10.1007/s11336-011-9207-7

DE LA TORRE, J. and DOUGLAS, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* **69** 333–353. MR2272454 https://doi.org/10.1007/BF02295640

Key words and phrases. Attribute hierarchy, Bayesian network, cognitive diagnostic model, directed graphical model, EM algorithm, identifiability.

GEORGE, A. C. and ROBITZSCH, A. (2015). Cognitive diagnosis models in R: A didactic. *Quant. Methods Psychol.* **11** 189–205.

GEORGE, A. C., ROBITZSCH, A., KIEFER, T., GROSS, J. and ÜNLÜ, A. (2016). The R package CDM for cognitive diagnosis models. *J. Stat. Softw.* **74** 1–24.

GIERL, M. J., LEIGHTON, J. P. and HUNKA, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about respondents' cognitive skills. In *Cognitive Diagnostic Assessment for Education: Theory and Applications* 242–274. Cambridge Univ. Press, Cambridge, UK.

GRÄTZER, G. (1971). *Lattice Theory. First Concepts and Distributive Lattices*. W. H. Freeman and Co., San Francisco, CA. MR0321817

GU, Y. and XU, G. (2019). Learning attribute patterns in high-dimensional structured latent attribute models. *J. Mach. Learn. Res.* **20** 1–58. MR3990469

GU, Y. and XU, G. (2023a). Identifiability of hierarchical latent attribute models. *Statist. Sinica* **33** 2561–2591. MR4647046

GU, Y. and XU, G. (2023b). A joint MLE approach to large-scale structured latent attribute analysis. *J. Amer. Statist. Assoc.* **118** 746–760. MR4571155 https://doi.org/10.1080/01621459.2021.1955689

HENSON, R. A., TEMPLIN, J. L. and WILLSE, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* **74** 191–210. MR2507377 https://doi.org/10.1007/s11336-008-9089-5

HO, N. and NGUYEN, X. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Ann. Statist.* **44** 2726–2755. MR3576559 https://doi.org/10.1214/16-AOS1444

HU, B. and TEMPLIN, J. (2020). Using diagnostic classification models to validate attribute hierarchies and evaluate model fit in Bayesian networks. *Multivar. Behav. Res.* **55** 300–311.

JUNKER, B. W. and SIJTSMA, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* **25** 258–272. MR1842982 https://doi.org/10.1177/01466210122032064

LEE, S. and GU, Y. (2024). Supplement to "Latent conjunctive Bayesian network: Unify attribute hierarchy and Bayesian network for cognitive diagnosis." https://doi.org/10.1214/23-AOAS1867SUPPA, https://doi.org/10.1214/23-AOAS1867SUPPB

LEIGHTON, J. and GIERL, M. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge Univ. Press, Cambridge.

LEIGHTON, J. P., GIERL, M. J. and HUNKA, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on tatsuoka's rule-space approach. *J. Educ. Meas.* **41** 205–237.

MA, C., OUYANG, J. and XU, G. (2023). Learning latent and hierarchical structures in Cognitive Diagnosis Models. *Psychometrika* **88** 175–207. MR4554878 https://doi.org/10.1007/s11336-022-09867-5

MULLIS, I. V., MARTIN, M. O., MINNICH, C. A., STANCO, G. M., ARORA, A., CENTURINO, V. A. and CASTLE, C. E. (2012). *TIMSS* 2011 *Encyclopedia: Education Policy and Curriculum in Mathematics and Science, Volume* 1: *AK*. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. The Morgan Kaufmann Series in Representation and Reasoning*. Morgan Kaufmann, San Mateo, CA. MR0965765

RUPP, A. A., TEMPLIN, J. and HENSON, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford, New York.

SHEN, X., PAN, W. and ZHU, Y. (2012). Likelihood-based selection and sharp parameter estimation. *J. Amer. Statist. Assoc.* **107** 223–232. MR2949354 https://doi.org/10.1080/01621459.2011.645783

SIMON, M. A. and TZUR, R. (2012). Explicating the role of mathematical tasks in conceptual learning: An elaboration of the hypothetical learning trajectory. **6** 91–104.

TANG, X., CHEN, Y., LI, X., LIU, J. and YING, Z. (2019). A reinforcement learning approach to personalized learning recommendation systems. *Br. J. Math. Stat. Psychol.* **72** 108–135.

TATSUOKA, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.* **20** 345–354.

TEMPLIN, J. and BRADSHAW, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika* **79** 317–339. MR3255122 https://doi.org/10.1007/s11336-013-9362-0

TEMPLIN, J. L. and HENSON, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* **11** 287.

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

VON DAVIER, M. (2008). A general diagnostic model applied to language testing data. *Br. J. Math. Stat. Psychol.* **61** 287–307. MR2649038 https://doi.org/10.1348/000711007X193957

VON DAVIER, M. and LEE, Y.-S. (2019). *Handbook of Diagnostic Classification Models*. Springer, Cham.

WANG, C. (2021). Using penalized EM algorithm to infer learning trajectories in latent transition CDM. *Psychometrika* **86** 167–189. MR4242639 https://doi.org/10.1007/s11336-020-09742-1

WANG, C. and GIERL, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *J. Educ. Meas.* **48** 165–187.

WANG, C. and LU, J. (2021). Learning attribute hierarchies from data: Two exploratory approaches. *J. Educ. Behav. Stat.* **46** 58–84.

XU, G. and SHANG, Z. (2018). Identifying latent structures in restricted latent class models. *J. Amer. Statist. Assoc.* **113** 1284–1295. MR3862357 https://doi.org/10.1080/01621459.2017.1340889

ZHAN, P., MA, W., JIAO, H. and DING, S. (2020). A sequential higher order latent structural model for hierarchical attributes in cognitive diagnostic assessments. *Appl. Psychol. Meas.* **44** 65–83.

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2012–2033
https://doi.org/10.1214/23-AOAS1868
© Institute of Mathematical Statistics, 2024

QUANTILE REGRESSION DECOMPOSITION ANALYSIS OF DISPARITY RESEARCH USING COMPLEX SURVEY DATA: APPLICATION TO DISPARITIES IN BMI AND TELOMERE LENGTH BETWEEN U.S. MINORITY AND WHITE POPULATION GROUPS

BY HYOKYOUNG G. HONG[1,a], BARRY I. GRAUBARD[1,b], JOSEPH L. GASTWIRTH[2,c] AND MI-OK KIM[3,d]

[1] *Biostatistics Branch, Division of Cancer Epidemiology and Genetics, NCI/NIH,* [a]*grace.hong@nih.gov,* [b]*graubarb@nih.gov*
[2] *Department of Statistics, George Washington University,* [c]*jlgast@gwu.edu*
[3] *Department of Epidemiology and Biostatistics, University of California San Francisco,* [d]*miok.kim@ucsf.edu*

We develop a quantile regression decomposition (QRD) method for analyzing observed disparities (OD) between population groups in socioeconomic and health-related outcomes for complex survey data. The conventional decomposition approaches use the conditional mean regression to decompose the disparity into two parts, the part explained by the difference arising from the different distributions in the explanatory covariates and the remaining part, which is unexplained by the covariates. Many socioeconomic and health outcomes exhibit heteroscedastic distributions, where the magnitude of observed disparities varies across different quantiles of these outcomes. Thus, differences in the explanatory covariates may account for varying differences in the OD across the quantiles of the outcome. The QRD can identify where there are greater differences in the outcome distribution, for example, 90th quantile, and how important the covariates are in explaining those differences. Much socioeconomic and health research relies on complex surveys, such as the National Health and Nutrition Examination Survey (NHANES), that oversample individuals from disadvantaged/minority population groups in order to provide improved precision. QRD has not been extended to the complex survey setting. We improve the QRD approach proposed in Machado and Mata (2005) to yield more reliable estimates at the quantiles, where the data are sparse, and extend it to the complex survey setting. We also propose a perturbation-based variance estimation method. Simulation studies indicate that the estimates of the unexplained portions of the OD across quantiles are unbiased and the coverage of the confidence intervals are close to nominal value. This methodology is used to study disparities in body mass index (BMI) and telomere length between race/ethnic groups estimated from the NHANES data.

REFERENCES

BELSON, W. A. (1956). A technique for studying the effects of a television broadcast. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **5** 195–202.

BLINDER, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *J. Hum. Resour.* **8** 436–455.

BLOOM, B. and BLACK, L. I. (2016). Health of non-Hispanic Asian adults: United States, 2010–2014. *NCHS Data Brief* 247 1–8.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2005). Health disparities experienced by black or African Americans-United States *Morb. Mort. Wkly. Rep.* **54** 1–3.

DRURY, S. S., ESTEVES, K., HATCH, V., WOODBURY, M., BORNE, S., ADAMSKI, A. and THEALL, K. P. (2015). Setting the trajectory: Racial disparities in newborn telomere length. *J. Pediatr.* **166** 1181–1186.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics* **38**. SIAM, Philadelphia, PA. MR0659849

Key words and phrases. Complex survey data, disparity decomposition, perturbation-based variance estimation, Peters–Belson, quantile regression.

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

FIRPO, S., FORTIN, N. M. and LEMIEUX, T. (2009). Unconditional quantile regressions. *Econometrica* **77** 953–973. MR2531365 https://doi.org/10.3982/ECTA6822

FORTIN, N., LEMIEUX, T. and FIRPO, S. (2011). Decomposition methods in economics. In *Handbook of Labor Economics* **4** 1–102. Elsevier, Amsterdam.

GASTWIRTH, J. L. and GREENHOUSE, S. W. (1995). Biostatistical concepts and methods in the legal setting. *Stat. Med.* **14** 1641–1653.

GERACI, M. (2016). Estimation of regression quantiles in complex surveys with data missing at random: An application to birthweight determinants. *Stat. Methods Med. Res.* **25** 1393–1421. MR3541104 https://doi.org/10.1177/0962280213484401

GRAUBARD, B. I., RAO, R. S. and GASTWIRTH, J. L. (2005). Using the Peters–Belson method to measure health care disparities from complex survey data. *Stat. Med.* **24** 2659–2668. MR2196206 https://doi.org/10.1002/sim.2135

HAYES, A., GEARON, E., BACKHOLER, K., BAUMAN, A. and PEETERS, A. (2015). Age-specific changes in BMI and BMI distribution among Australian adults using cross-sectional surveys from 1980 to 2008. *Int. J. Obes.* **39** 1209–1216.

HONG, H. G, GRAUBARD, B. I, GASTWIRTH, J. L and KIM, M.-O (2024). Supplement to “Quantile regression decomposition analysis of disparity research using complex survey data: Application to disparities in BMI and telomere length between U.S. minority and white population groups.” https://doi.org/10.1214/23-AOAS1868SUPP

HUNT, S. C., CHEN, W., GARDNER, J. P., KIMURA, M., SRINIVASAN, S. R., ECKFELDT, J. H., BERENSON, G. S. and AVIV, A. (2008). Leukocyte telomeres are longer in African Americans than in whites: The national heart, lung, and blood institute family heart study and the bogalusa heart study. *Aging Cell* **7** 451–458.

JACKSON, J. W. (2021). Meaningful causal decompositions in health equity research: Definition, identification, and estimation through a weighting framework. *Epidemiology* **32** 282.

JACKSON, J. W. and VANDERWEELE, T. J. (2018). Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology* **29** 825.

JEFFRIES, N., ZASLAVSKY, A. M., DIEZ ROUX, A. V., CRESWELL, J. W., PALMER, R. C., GREGORICH, S. E., RESCHOVSKY, J. D., GRAUBARD, B. I., CHOI, K. et al. (2019). Methodological approaches to understanding causes of health disparities. *Amer. J. Publ. Health* **109** S28–S33.

JIN, Z., YING, Z. and WEI, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88** 381–390. MR1844838 https://doi.org/10.1093/biomet/88.2.381

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

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

KOLENIKOV, S. (2010). Resampling variance estimation for complex survey data. *Stata J.* **10** 165–199.

KORN, E. L. and GRAUBARD, B. I. (2011). *Analysis of Health Surveys*. Wiley, NJ.

LI, Y., GRAUBARD, B. I., HUANG, P. and GASTWIRTH, J. L. (2015). Extension of the Peters–Belson method to estimate health disparities among multiple groups using logistic regression with survey data. *Stat. Med.* **34** 595–612. MR3301599 https://doi.org/10.1002/sim.6357

LUCAS, J. W., FREEMAN, G. and ADAMS, P. F. (2016). *Health of Hispanic Adults*: *United States*, 2010–2014. National Center for Health Statistics, Hyattsville, MD.

MACHADO, J. A. F. and MATA, J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *J. Appl. Econometrics* **20** 445–465. MR2143445 https://doi.org/10.1002/jae.788

MELLY, B. (2005). Decomposition of differences in distribution using quantile regression. *Labour Econ.* **12** 577–590.

NATIONAL CENTER FOR HEALTH STATISTICS (2018). *National Health and Nutrition Examination Survey*: *Analytic Guidelines*. 2011–2014 *and* 2015–2016, Centers for Disease Control and Prevention, Atlanta, GA.

NATIONAL CENTER FOR HEALTH STATISTICS (2022). *National health and nutrition examination survey data*. Centers for Disease Control and Prevention (CDC), Hyattsville, MD. U.S. Department of Health and Human Services, centers for disease control and prevention. https://www.cdc.gov/nchs/nhanes/index.htm.

NEEDHAM, B. L., ADLER, N., GREGORICH, S., REHKOPF, D., LIN, J., BLACKBURN, E. H. and EPEL, E. S. (2013). Socioeconomic status, health behavior, and leukocyte telomere length in the national health and nutrition examination survey, 1999–2002. *Soc. Sci. Med.* **85** 1–8.

NEEDHAM, B. L., SALERNO, S., ROBERTS, E., BOSS, J., ALLGOOD, K. L. and MUKHERJEE, B. (2019). Do black/white differences in telomere length depend on socioeconomic status? *Biodemogr. Soc. Biol.* **65** 287–312.

OAXACA, R. (1973). Male-female wage differentials in urban labor markets. *Internat. Econom. Rev.* **14** 693–709.

PARK, S. and HE, X. (2017). Hypothesis testing for regional quantiles. *J. Statist. Plann. Inference* **191** 13–24. MR3679106 https://doi.org/10.1016/j.jspi.2017.06.002

PARK, S., LEE, E. R. and HONG, H. G. (2023). Varying-coefficients for regional quantile via KNN-based LASSO with applications to health outcome study. *Stat. Med.* **42** 3903–3918. MR4637294 https://doi.org/10.1002/sim.9839

PETERS, C. C. (1941). A method of matching groups for experiment with no loss of population. *J. Educ. Res.* **34** 606–612.

R CORE TEAM (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.

REWAK, M., BUKA, S., PRESCOTT, J., VIVO, I. D., LOUCKS, E. B., KAWACHI, I., NON, A. L. and KUBZANSKY, L. D. (2014). Race-related health disparities and biological aging: Does rate of telomere shortening differ across blacks and whites? *Biol. Psychol.* **99** 92–99. https://doi.org/10.1016/j.biopsycho.2014.03.007

RUST, K. F. and RAO, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Stat. Methods Med. Res.* **5** 283–310.

SINCLAIR, M. D. and PAN, Q. (2009). Using the Peters–Belson method in equal employment opportunity personnel evaluations. *Law Probab. Risk* **8** 95–117.

VAISERMAN, A. and KRASNIENKOV, D. (2020). Telomere length as a marker of biological age: State-of-the-art, open issues, and future perspectives. *Front. Genet.* **11** 630186. https://doi.org/10.3389/fgene.2020.630186

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
2024, Vol. 18, No. 3, 2034–2061
https://doi.org/10.1214/23-AOAS1869

BAYESIAN HIDDEN MARKOV MODELS FOR LATENT VARIABLE LABELING ASSIGNMENTS IN CONFLICT RESEARCH: APPLICATION TO THE ROLE CEASEFIRES PLAY IN CONFLICT DYNAMICS

BY JONATHAN P. WILLIAMS[1,a], GUDMUND H. HERMANSEN[2,b], HÅVARD STRAND[2,d], GOVINDA CLAYTON[3,c] AND HÅVARD MOKLEIV NYGÅRD[4,e]

[1]*Department of Statistics, North Carolina State University,* [a]*jwilli27@ncsu.edu*
[2]*Department of Mathematics, University of Oslo,* [b]*gudmunhh@math.uio.no*
[3]*Center for Security Studies, ETH Zürich,* [c]*govinda.clayton@sipo.gess.ethz.ch*
[4]*Department of Peace and Conflict Dynamics, Peace Research Institute Oslo (PRIO),* [d]*hs@prio.org,* [e]*havnyg@prio.org*

A crucial challenge for solving problems in conflict research is in leveraging the semisupervised nature of the data that arise. Observed response data, such as counts of battle deaths over time, indicate latent processes of interest, such as intensity and duration of conflicts, but defining and labeling instances of these unobserved processes requires nuance and imprecision. The availability of such labels, however, would make it possible to study the effect of intervention-related predictors—such as ceasefires—directly on conflict dynamics (e.g., latent intensity) rather than through an intermediate proxy, like observed counts of battle deaths. Motivated by this problem and the new availability of the ETH-PRIO Civil Conflict Ceasefires data set, we propose a Bayesian autoregressive (AR) hidden Markov model (HMM) framework as a sufficiently flexible machine learning approach for semisupervised regime labeling with uncertainty quantification. We motivate our approach by illustrating the way it can be used to study the role that ceasefires play in shaping conflict dynamics. This ceasefires data set is the first systematic and globally comprehensive data on ceasefires, and our work is the first to analyze this new data and to explore the effect of ceasefires on conflict dynamics in a comprehensive and cross-country manner.

REFERENCES

ÅKEBO, M. (2016). *Ceasefire Agreements and Peace Processes*: *A Comparative Study*. Routledge, London.

AARY, V. (1995). Concluding hostilities: Humanitarian provisions in cease-fire agreements. *Mil. Law Rev.* **148** 186–273.

AKEBO, M. (2016). *Ceasefire Agreements and Peace Processes*: *A Comparative Study*. Taylor & Francis, London.

ANDERS, T. (2020). Territorial control in civil wars: Theory and measurement using machine learning. *J. Peace Res.* **57** 701–714.

BARA, C., CLAYTON, G. and RUSTAD, S. A. (2021). Understanding ceasefires. *Int. Peacekeep.* **28** 329–340.

BESLEY, T., FETZER, T. and MUELLER, H. (2021). How big is the media multiplier? Evidence from dyadic news data. CESifo Working Paper.

BRICKHILL, J. (2018). *Mediating Security Arrangements in Peace Processes*: *Critical Perspectives from the Field*. ETH Press, Zurich. OCLC: 1043551307.

BRUNBORG, H., LYNGSTAD, T. H. and URDAL, H. (2003). Accounting for genocide: How many were killed in Srebrenica? *Eur. J. Popul.* **19** 229–248. https://doi.org/10.1023/A:1024949307841

BUCHANAN, C., CLAYTON, G. and RAMSBOTHAM, A. (2021). *Ceasefire Monitoring*: *Developments and Complexities*. Accord Spotlight: Conciliation Resources, London, UK.

CHEN, B., SHRIVASTAVA, A. and STEORTS, R. C. (2018). Unique entity estimation with application to the Syrian conflict. *Ann. Appl. Stat.* **12** 1039–1067. MR3834294 https://doi.org/10.1214/18-AOAS1163

CHOUNET-CAMBAS, L. (2011). *Negotiating Ceasefires*. *Mediation Practice Series*. Centre for Humanitarian Dialogue, Geneva.

Key words and phrases. State space model, multistate model, discrete-time Markov process, discrete-valued time series, count-valued time series.

CLAYTON, G., MASON, S., STICHER, V. and WIEHLER, C. (2019). Ceasefires in intra-state peace processes. *CSS Anal. Secur. Policy* **252**.

CLAYTON, G., NATHAN, L. and WIEHLER, C. (2021). Ceasefire success: A conceptual framework. *Int. Peacekeep.* **28** 341–365.

CLAYTON, G., NYGÅRD, H. M., RUSTAD, S. A. and STRAND, H. (2023). Ceasefires in civil conflict: A research agenda. *J. Confl. Resolut.* **67** 1279–1295. https://doi.org/10.1177/00220027221128300

CLAYTON, G. et al. (2023). Ceasefires in civil conflict: A research agenda. *J. Confl. Resolut.* **67** 1430–1451.

CLAYTON, G. and STICHER, V. (2021). The logic of ceasefires in civil war. *Int. Stud. Q.* Online first.

COPPEDGE, M., GERRING, J., KNUTSEN, C. H., LINDBERG, S. I., TEORELL, J., ALTMAN, D., BERNHARD, M., FISH, M. S., GLYNN, A. et al. (2019). V-Dem [Country-Year/Country-Date] Dataset v9.

CRISMAN-COX, C. (2022). Democracy, reputation for resolve, and civil conflict. *J. Peace Res.* **59** 382–394. https://doi.org/10.1177/00223433211024697

DAHL, R. A. (1971). *Polyarchy: Political Participation and Opposition*. Yale Univ. Press, New Haven, CT.

DAVENPORT, C., NYGÅRD, H. M., FJELDE, H. and ARMSTRONG, D. (2019). The consequences of contention: Understanding the aftereffects of political conflict and violence. *Annu. Rev. Pol. Sci.* **22** 1–30.

DAVIS, R. A., FOKIANOS, K., HOLAN, S. H., JOE, H., LIVSEY, J., LUND, R., PIPIRAS, V. and RAVISHANKER, N. (2021). Count time series: A methodological review. *J. Amer. Statist. Assoc.* **116** 1533–1547. MR4309291 https://doi.org/10.1080/01621459.2021.1904957

DAWKINS, S. (2021). The problem of the missing dead. *J. Peace Res.* **58** 1098–1116. https://doi.org/10.1177/0022343320962159

DE SOTO, A. (1999). Ending violent conflict in El Salvador. In *Herding Cats: Multiparty Mediation in a Complex World* United States Institute of Peace Press, Washington D.C.

DUKALSKIS, A. (2015). Why do some insurgent groups agree to cease-fires while others do not? A within-case analysis of Burma/Myanmar, 1948–2011. *Stud. Confl. Terrorism* **38** 841–863. https://doi.org/10.1080/1057610X.2015.1056631

FAZAL, T. M. (2014). Dead wrong?: Battle deaths, military medicine, and exaggerated reports of war's demise. *Int. Secur.* **39** 95–125.

FORTNA, V. P. (2003). Scraps of paper? Agreements and the durability of peace. *Int. Organ.* **57** 337–372.

FORTNA, V. P. (2004). *Peace Time: Cease-Fire Agreements and the Durability of Peace*. Princeton Univ. Press, Princeton.

GEORGE, A. L. and BENNET, A. (2005). *Case Studies and Theory Development in the Social Sciences*. Cambridge Univ. Press, Cambridge.

GHOBARAH, H. A., HUTH, P. and RUSSETT, B. (2003). Civil wars kill and maim people—Long after the shooting stops. *Amer. Polit. Sci. Rev.* **97** 189–202.

GLEDITSCH, K. S. (2007). Transnational dimensions of civil war. *J. Peace Res.* **44** 293–309. https://doi.org/10.1177/0022343307076637

GLEDITSCH, N. P., WALLENSTEEN, P., ERIKSSON, M., SOLLENBERG, M. and STRAND, H. (2002). Armed conflict 1946–2001: A new dataset. *J. Peace Res.* **39** 615–637.

HANSON, K. (2021). Live and let live: Explaining long-term truces in separatist conflicts. *Int. Peacekeep.* **28** 393–415.

HEGRE, H., ELLINGSEN, T., GATES, S. and GLEDITSCH, N. P. (2001). Toward a democratic civil peace? Democracy, political change, and civil war, 1816–1992. *Amer. Polit. Sci. Rev.* **95** 33–48. https://doi.org/10.1017/S0003055401000119

HÖGBLADH, S. (2023). UCDP GED Codebook version 23.1. Department of Peace and Conflict Research, Uppsala University.

HÖGLUND, K. (2005). Violence and the peace process in Sri Lanka. *Civil Wars* **7** 156–170.

HÖGLUND, K. (2011). Tactics in negotiations between states and extremists: The role of cease-fires and counterterrorist measures. In *Engaging Extremists: Trade-Offs, Timing, and Diplomacy* (I. W. Zartman and G. O. Faure, eds.) United States Institute of Peace, Washington D.C.

HOLTERMANN, H. (2021). Blinding the elephant: Combat, information, and rebel violence. *Terrorism Polit. Violence* **33** 1469–1491.

JAKOBSEN, J. H. (2021). Application of count time series to battle deaths. Master's thesis, Univ. Oslo. Available at http://urn.nb.no/URN:NBN:no-90532.

JARMAN, N. (2004). From war to peace? Changing patterns of violence in northern Ireland, 1990–2003. *Terrorism Polit. Violence* **16** 420–438.

KARAKUS, D. C. and SVENSSON, I. (2020). Between the bombs: Exploring partial ceasefires in the Syrian civil war, 2011–2017. *Terrorism Polit. Violence* **32** 681–700. https://doi.org/10.1080/09546553.2017.1393416

KOLÅS, Å. (2011). Naga militancy and violent politics in the shadow of ceasefire. *J. Peace Res.* **48** 781–792.

KREUTZ, J. (2010). How and when armed conflicts end: Introducing the UCDP conflict termination dataset. *J. Peace Res.* **47** 243–250.

KRTSCH, R. (2021). The tactical use of civil resistance by rebel groups: Evidence from India's Maoist insurgency. *J. Confl. Resolut.* **65** 1251–1277. https://doi.org/10.1177/0022002721995547

LACINA, B. (2006). Explaining the severity of civil wars. *J. Confl. Resolut.* **50** 276–289.

LACINA, B. and GLEDITSCH, N. P. (2012). The waning of war is real: A response to gohdes and price. *J. Confl. Resolut.* **57** 1109–1127.

LUTSCHER, P. M., WEIDMANN, N. B., ROBERTS, M. E., JONKER, M., KING, A. and DAINOTTI, A. (2020). At home and abroad: The use of denial-of-service attacks during elections in nondemocratic regimes. *J. Confl. Resolut.* **64** 373–401. https://doi.org/10.1177/0022002719861676

LUTTWAK, E. N. (1999). Give war a chance. *Foreign Aff.* 36–44.

MAHIEU, S. (2007). When should mediators interrupt a civil war? The best timing for a ceasefire. *Int. Negot.* **12** 207–228.

PALIK, J. (2021). Watchdogs of pause: The challenges of ceasefire monitoring in Yemen. *Int. Peacekeep.* **28** 444–469. https://doi.org/10.1080/13533312.2021.1918004

PETROFF, V. B., BOND, J. H. and BOND, D. H. (2013). Using hidden Markov models to predict terror before it hits (again). In *Handbook of Computational Approaches to Counterterrorism* 163–180. Springer, Berlin.

PINAUD, M. (2021). Home-grown peace: Civil society roles in ceasefire monitoring. *Int. Peacekeep.* **28** 470–495. https://doi.org/10.1080/13533312.2020.1861943

PLANK, F. (2017). When peace leads to divorce: The splintering of rebel groups in powersharing agreements. *Civil Wars* **19** 176–197. https://doi.org/10.1080/13698249.2017.1372004

PRICE, M. and BALL, P. (2014). Big data, selection bias, and the statistical patterns of mortality in conflict. *SAIS Rev. Int. Aff.* **34** 9–20.

RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77** 257–286.

RALEIGH, C., KISHI, R. and LINKE, A. (2023). Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanit. Soc. Sci. Commun.* **10** 1–17.

RANDAHL, D. and VEGELIUS, J. (2022). Predicting escalating and de-escalating violence in Africa using Markov models. *Int. Interact.* **48** 597–613.

REEDER, B. W. and SEEBERG, M. B. (2018). Fighting your friends? A study of intra-party violence in sub-Saharan Africa. *Democratization* **25** 1033–1051. https://doi.org/10.1080/13510347.2018.1441291

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880. MR0770281

SADINLE, M. (2014). Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *Ann. Appl. Stat.* **8** 2404–2434. MR3292503 https://doi.org/10.1214/14-AOAS779

SADINLE, M. (2018). Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. *Ann. Appl. Stat.* **12** 1013–1038. MR3834293 https://doi.org/10.1214/18-AOAS1178

SATTEN, G. A. and LONGINI JR, I. M. (1996). Markov chains with measurement error: Estimating the 'true'course of a marker of the progression of human immunodeficiency virus disease. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **45** 275–295.

SCHRODT, P. A. (1997a). Pattern recognition of international crises using hidden Markov models. *Nonlinear Models Methods Polit. Sci.*

SCHRODT, P. A. (1997b). Early warning of conflict in southern Lebanon using hidden Markov models. In *American Political Science Association*.

SCHRODT, P. A. (2006). Forecasting conflict in the Balkans using hidden Markov models. In *Programming for Peace* 161–184. Springer, Berlin.

SMITH, J. D. D. (1995). *Stopping Wars: Defining the Obstacles to Cease-Fire*. Westview Press, Boulder, CO.

SOSNOWSKI, M. (2020). Ceasefires as violent state-building: Local truce and reconciliation agreements in the Syrian civil war. *Conflict, Security, Development* **20** 273–292. https://doi.org/10.1080/14678802.2019.1679561

SUNDBERG, R. and MELANDER, E. (2013). Introducing the UCDP georeferenced event dataset. *J. Peace Res.* **50** 523–532. https://doi.org/10.1177/0022343313484347

TAI, X. H., MEHRA, S. and BLUMENSTOCK, J. E. (2022). Mobile phone data reveal the effects of violence on internal displacement in Afghanistan. *Nat. Hum. Behav.* **6** 624–634. https://doi.org/10.1038/s41562-022-01336-4

WATERMAN, A. (2021). Ceasefires and state order-making in Naga Northeast India. *Int. Peacekeep.* **28** 496–525. https://doi.org/10.1080/13533312.2020.1821365

WEIDMANN, N. B. (2015). On the accuracy of media-based conflict event data. *J. Confl. Resolut.* **59** 1129–1149. https://doi.org/10.1177/0022002714530431

WEISS, C. H. (2018). *An Introduction to Discrete-Valued Time Series*. Wiley, New York.

WILLIAMS, J. P., HERMANSEN, G. H., STRAND, H., CLAYTON, G. and NYGÅRD, H. M. (2024). Supplement to "Bayesian hidden Markov models for latent variable labeling assignments in conflict research: Application to the role ceasefires play in conflict dynamics." https://doi.org/10.1214/23-AOAS1869SUPP

WILLIAMS, J. P., STORLIE, C. B., THERNEAU, T. M., JACK, C. R. JR. and HANNIG, J. (2020). A Bayesian approach to multistate hidden Markov models: Application to dementia progression. *J. Amer. Statist. Assoc.* **115** 16–31. MR4078442 https://doi.org/10.1080/01621459.2019.1594831

WOOD, R. M. (2014). From loss to looting? Battlefield costs and rebel incentives for violence. *Int. Organ.* **68** 979–999. https://doi.org/10.1017/S0020818314000204

WOODS, K. (2011). Ceasefire capitalism: Military–private partnerships, resource concessions and military–state building in the Burma–China borderlands. *J. Peasant Stud.* **38** 747–770. https://doi.org/10.1080/03066150.2011.607699

XU, H.-Y., XIE, M., GOH, T. N. and FU, X. (2012). A model for integer-valued time series with conditional overdispersion. *Comput. Statist. Data Anal.* **56** 4229–4242. MR2957867 https://doi.org/10.1016/j.csda.2012.04.011

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2062–2079
https://doi.org/10.1214/23-AOAS1870
© Institute of Mathematical Statistics, 2024

SEMIPARAMETRIC ESTIMATION FOR DYNAMIC NETWORKS WITH SHIFTED CONNECTING INTENSITIES

BY ZITONG ZHANG[a] AND SHIZHE CHEN[b]

Department of Statistics, University of California Davis, [a]*zztzhang@ucdavis.edu,* [b]*szdchen@ucdavis.edu*

Neural circuits are of paramount importance in the nervous system, as they are the essential infrastructure in guiding animal behavior. However, modeling the development of neural circuits poses significant challenges due to inherent properties of the development process. First, the neural circuit development process is transient, where the course of development can only be observed once. Second, despite potentially sharing similar underlying mechanisms for development, neural circuits from different subjects possess distinct sets of neurons, which limits the sharing of information across subjects. Third, neurons have diverse, unobserved activation times, which may obscure the analysis of neural activities. In light of these challenges, this study presents a novel approach aimed at clustering neurons based on their connecting behaviors while accommodating disparities at the neuron level. To this end, we propose a dynamic stochastic block model that accommodates unknown time shifts. We establish the conditions that guarantee the identifiability of cluster memberships of nodes and representative connecting intensities across clusters. Using methods for shape invariant models, we propose computationally efficient semiparametric estimation procedures to simultaneously estimate time shifts, cluster memberships, and connecting intensities. We illustrate the performance of the proposed procedures via extensive simulation experiments. We further apply the proposed method on a motor circuit development data from zebrafish to reveal distinct roles of neurons and identify representative connecting behaviors.

REFERENCES

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. https://doi.org/10.1109/34.865189

BIGOT, J. and GADAT, S. (2010). A deconvolution approach to estimation of a common shape in a shifted curves model. *Ann. Statist.* **38** 2422–2464. MR2676894 https://doi.org/10.1214/10-AOS800

BIGOT, J., GADAT, S., KLEIN, T. and MARTEAU, C. (2013). Intensity estimation of non-homogeneous Poisson processes from shifted trajectories. *Electron. J. Stat.* **7** 881–931. MR3044503 https://doi.org/10.1214/13-EJS794

BIGOT, J. and GENDRE, X. (2013). Minimax properties of Fréchet means of discretely sampled curves. *Ann. Statist.* **41** 923–956. MR3099126 https://doi.org/10.1214/13-AOS1104

BLANKENSHIP, A. G. and FELLER, M. B. (2010). Mechanisms underlying spontaneous patterned activity in developing neural circuits. *Nat. Rev. Neurosci.* **11** 18–29. https://doi.org/10.1038/nrn2759

BONTEMPS, D. and GADAT, S. (2014). Bayesian methods for the shape invariant model. *Electron. J. Stat.* **8** 1522–1568. MR3263130 https://doi.org/10.1214/14-EJS933

DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes. Vol. I: Elementary Theory and Methods*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR1950431

DAUDIN, J.-J., PICARD, F. and ROBIN, S. (2008). A mixture model for random graphs. *Stat. Comput.* **18** 173–183. MR2390817 https://doi.org/10.1007/s11222-007-9046-7

DIESNER, J. and CARLEY, K. M. (2005). Exploration of communication networks from the Enron email corpus. In *SIAM International Conference on Data Mining*: *Workshop on Link Analysis*, 3–14. Counterterrorism and Security, Newport Beach, CA.

Key words and phrases. Dynamic networks, semiparametric estimation, shape invariant models, stochastic block models.

GAO, C., LU, Y. and ZHOU, H. H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43** 2624–2652. MR3405606 https://doi.org/10.1214/15-AOS1354

GIORGI, D., MATIAS, C., REBAFKA, T. and VILLERS, F. (2018). ppsbm: Clustering in longitudinal networks R package version 0.2.2.

HAYTHORNTHWAITE, C. (1996). Social network analysis: An approach and technique for the study of information exchange. *Libr. Inf. Sci. Res.* **18** 323–342.

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

JAO, L.-E., APPEL, B. and WENTE, S. R. (2012). A zebrafish model of lethal congenital contracture syndrome 1 reveals Gle1 function in spinal neural precursor survival and motor axon arborization. *Development* **139** 1316–1326. https://doi.org/10.1242/dev.074344

KELLER, P. (2019). Longitudinal functional imaging data for the zebrafish embryonic spinal cord. Janelia Research Campus. Dataset. Available at https://doi.org/10.25378/janelia.7605824.v1.

KLEINBAUM, D. G. and KLEIN, M. (2012). *Survival Analysis*: *A Self-Learning Text*, 3rd ed. *Statistics for Biology and Health*. Springer, New York. MR2882858 https://doi.org/10.1007/978-1-4419-6646-9

KNEIP, A. and ENGEL, J. (1995). Model estimation in nonlinear regression under shape invariance. *Ann. Statist.* **23** 551–570. MR1332581 https://doi.org/10.1214/aos/1176324535

KREISS, A., MAMMEN, E. and POLONIK, W. (2019). Nonparametric inference for continuous-time event counting and link-based dynamic network models. *Electron. J. Stat.* **13** 2764–2829. MR3995010 https://doi.org/10.1214/19-EJS1588

KRIVITSKY, P. N. and HANDCOCK, M. S. (2014). A separable model for dynamic networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 29–46. MR3153932 https://doi.org/10.1111/rssb.12014

LEI, J., CHEN, K. and LYNCH, B. (2020). Consistent community detection in multi-layer network data. *Biometrika* **107** 61–73. MR4064140 https://doi.org/10.1093/biomet/asz068

LIN, L., HE, Z. and PEETA, S. (2018). Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. *Transp. Res., Part C, Emerg. Technol.* **97** 258–276.

LONGEPIERRE, L. and MATIAS, C. (2019). Consistency of the maximum likelihood and variational estimators in a dynamic stochastic block model. *Electron. J. Stat.* **13** 4157–4223. MR4021264 https://doi.org/10.1214/19-EJS1624

LOUPOS, P., NATHAN, A. and CERF, M. (2019). Starting cold: The power of social networks in predicting non-contractual customer behavior. Available at SSRN 3001978.

MATIAS, C. and MIELE, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1119–1141. MR3689311 https://doi.org/10.1111/rssb.12200

MATIAS, C., REBAFKA, T. and VILLERS, F. (2018). A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika* **105** 665–680. MR3842891 https://doi.org/10.1093/biomet/asy016

MENELAOU, E. and MCLEAN, D. L. (2019). Hierarchical control of locomotion by distinct types of spinal V2a interneurons in zebrafish. *Nat. Commun.* **10** 4197. https://doi.org/10.1038/s41467-019-12240-3

NISHIMARU, H., RESTREPO, C. E., RYGE, J., YANAGAWA, Y. and KIEHN, O. (2005). Mammalian motor neurons corelease glutamate and acetylcholine at central synapses. *Proc. Natl. Acad. Sci. USA* **102** 5245–5249.

OSATUYI, B. (2013). Information sharing on social media sites. *Comput. Hum. Behav.* **29** 2622–2631.

PAUL, S. and CHEN, Y. (2020). A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging. *Ann. Appl. Stat.* **14** 993–1029. MR4117838 https://doi.org/10.1214/20-AOAS1339

PAVLOVIĆ, D. M., GUILLAUME, B. R. L., TOWLSON, E. K., KUEK, N. M. Y., AFYOUNI, S., VÉRTES, P. E., YEO, B. T. T., BULLMORE, E. T. and NICHOLS, T. E. (2020). Multi-subject stochastic blockmodels for adaptive analysis of individual differences in human brain network cluster structure. *NeuroImage* **220** 116611. https://doi.org/10.1016/j.neuroimage.2020.116611

PENSKY, M. (2019). Dynamic network models and graphon estimation. *Ann. Statist.* **47** 2378–2403. MR3953455 https://doi.org/10.1214/18-AOS1751

SILVERMAN, B. W. (2017). *Density Estimation for Statistics and Data Analysis*. Routledge, New York. https://doi.org/10.1201/9781315140919

SONG, J., PALLUCCHI, I., AUSBORN, J., AMPATZIS, K., BERTUZZI, M., FONTANEL, P., PICTON, L. D. and EL MANIRA, A. (2020). Multiple rhythm-generating circuits act in tandem with pacemaker properties to control the start and speed of locomotion. *Neuron* **105** 1048–1061.

VIMOND, M. (2010). Efficient estimation for a subclass of shape invariant models. *Ann. Statist.* **38** 1885–1912. MR2662362 https://doi.org/10.1214/07-AOS566

WAN, Y., WEI, Z., LOOGER, L. L., KOYAMA, M., DRUCKMANN, S. and KELLER, P. J. (2019). Single-cell reconstruction of emerging population activity in an entire developing circuit. *Cell* **179** 355–372.e23. https://doi.org/10.1016/j.cell.2019.08.039

WENNER, P. and O'DONOVAN, M. J. (2001). Mechanisms that initiate spontaneous network activity in the developing chick spinal cord. *J. Neurophysiol.* **86** 1481–1498. https://doi.org/10.1152/jn.2001.86.3.1481

XING, E. P., FU, W. and SONG, L. (2010). A state-space mixed membership blockmodel for dynamic network tomography. *Ann. Appl. Stat.* **4** 535–566. MR2758639 https://doi.org/10.1214/09-AOAS311

XU, K. S. (2015). Stochastic block transition models for dynamic networks. *J. Mach. Learn. Res.* **38** 1079–1087.

XU, K. S. and HERO, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE J. Sel. Top. Signal Process.* **8** 552–562. https://doi.org/10.1109/JSTSP.2014.2310294

YANG, T., CHI, Y., ZHU, S., GONG, Y. and JIN, R. (2011). Detecting communities and their evolutions in dynamic social networks—a Bayesian approach. *Mach. Learn.* **82** 157–189. MR3108191 https://doi.org/10.1007/s10994-010-5214-7

ZHANG, J., SUN, W. W. and LI, L. (2020). Mixed-effect time-varying network model and application in brain connectivity analysis. *J. Amer. Statist. Assoc.* **115** 2022–2036. MR4189774 https://doi.org/10.1080/01621459.2019.1677242

ZHANG, Z. and CHEN, S. (2024). Supplement to "Semiparametric estimation for dynamic networks with shifted connecting intensities." https://doi.org/10.1214/23-AOAS1870SUPPA, https://doi.org/10.1214/23-AOAS1870SUPPB

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2080–2095
https://doi.org/10.1214/23-AOAS1871

A NONPARAMETRIC MIXED-EFFECTS MIXTURE MODEL FOR PATTERNS OF CLINICAL MEASUREMENTS ASSOCIATED WITH COVID-19

BY XIAORAN MA[1,a], WENSHENG GUO[2,d], MENGYANG GU[1,b], LEN USVYAT[3,e], PETER KOTANKO[4,f] AND YUEDONG WANG[1,c]

[1]*Department of Statistics and Applied Probability, University of California, Santa Barbara,* [a]*xiaoran_ma@pstat.ucsb.edu,* [b]*mengyang@pstat.ucsb.edu,* [c]*yuedong@pstat.ucsb.edu*

[2]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania,* [d]*wguo@upenn.edu*

[3]*Fresenius Medical Care,* [e]*Len.Usvyat@freseniusmedicalcare.com*

[4]*Renal Research Institute,* [f]*Peter.Kotanko@rriny.com*

Some patients with COVID-19 show changes in signs and symptoms, such as temperature and oxygen saturation days before being positively tested for SARS-CoV-2, while others remain asymptomatic. It is important to identify these subgroups and to understand what biological and clinical predictors are related to these subgroups. This information will provide insights into how the immune system may respond differently to infection and can further be used to identify infected individuals. We propose a flexible nonparametric mixed-effects mixture model that identifies risk factors and classifies patients with biological changes. We model the latent probability of biological changes using a logistic regression model and trajectories in the latent groups using smoothing splines. We developed an EM algorithm to maximize the penalized likelihood for estimating all parameters and mean functions. We evaluate our methods by simulations and apply the proposed model to investigate changes in temperature in a cohort of COVID-19-infected hemodialysis patients.

REFERENCES

ARAGAM, B., DAN, C., XING, E. P. and RAVIKUMAR, P. (2020). Identifiability of nonparametric mixture models and Bayes optimal clustering. *Ann. Statist.* **48** 2277–2302. MR4134795 https://doi.org/10.1214/19-AOS1887

BHAVANI, S. V., WILEY, Z., VERHOEF, P. A., COOPERSMITH, C. M. and OFOTOKUN, I. (2022). Racial differences in detection of fever using temporal vs oral temperature measurements in hospitalized patients. *JAMA* **328** 885. https://doi.org/10.1001/jama.2022.12290

BIVONA, G., AGNELLO, L. and CIACCIO, M. (2021). Biomarkers for prognosis and treatment response in COVID-19 patients. *Ann. Lab. Med.* **41** 540–548. https://doi.org/10.3343/alm.2021.41.6.540

BOUVEYRON, C. and BRUNET-SAUMARD, C. (2014). Model-based clustering of high-dimensional data: A review. *Comput. Statist. Data Anal.* **71** 52–78. MR3131954 https://doi.org/10.1016/j.csda.2012.12.008

BYRD, R. H., LU, P., NOCEDAL, J. and ZHU, C. Y. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **16** 1190–1208. MR1346301 https://doi.org/10.1137/0916069

CHAUDHURI, S., LASKY, R., JIAO, Y., LARKIN, J., MONAGHAN, C., WINTER, A., NERI, L., KOTANKO, P., HYMES, J. et al. (2022). Trajectories of clinical and laboratory characteristics associated with COVID-19 in hemodialysis patients by survival. *Hemodial. Int.* **26** 94–107. https://doi.org/10.1111/hdi.12977

CHETVERIKOV, D., LIAO, Z. and CHERNOZHUKOV, V. (2021). On cross-validated Lasso in high dimensions. *Ann. Statist.* **49** 1300–1317. MR4298865 https://doi.org/10.1214/20-aos2000

DA ROSA MESQUITA, R., FRANCELINO SILVA JUNIOR, L. C., SANTOS SANTANA, F. M., FARIAS DE OLIVEIRA, T., CAMPOS ALCÂNTARA, R., MONTEIRO ARNOZO, G., RODRIGUES DA SILVA FILHO, E., GALDINO DOS SANTOS, A. G., OLIVEIRA DA CUNHA, E. J. et al. (2021). Clinical manifestations of COVID-19 in the general population: Systematic review. *Wien. Klin. Wochenschr.* **133** 377–382. https://doi.org/10.1007/s00508-020-01760-4

Key words and phrases. Clustering, Coronavirus Disease 2019, COVID-19, EM algorithm, mixed-effects model, mixture model, SARS-CoV-2, severe acute respiratory syndrome coronavirus 2, spline.

DE MORAES BATISTA, A. F., MIRAGLIA, J. L., RIZZI DONATO, T. H. and PORTO CHIAVEGATTO FILHO, A. D. (2020). COVID-19 diagnosis prediction in emergency care patients: A machine learning approach. Preprint. Epidemiology. Available at https://doi.org/10.1101/2020.04.04.20052092.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22. https://doi.org/10.18637/jss.v033.i01

FRÜHWIRTH-SCHNATTER, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J. Amer. Statist. Assoc.* **96** 194–209. MR1952732 https://doi.org/10.1198/016214501750333063

GALLO MARIN, B., AGHAGOLI, G., LAVINE, K., YANG, L., SIFF, E. J., CHIANG, S. S., SALAZAR-MATHER, T. P., DUMENCO, L., SAVARIA, M. C. et al. (2021). Predictors of COVID-19 severity: A literature review. *Rev. Med. Virol.* **31** 1–10. https://doi.org/10.1002/rmv.2146

GU, C. (2013). *Smoothing Spline ANOVA Models*, 2nd ed. *Springer Series in Statistics* **297**. Springer, New York. MR3025869 https://doi.org/10.1007/978-1-4614-5369-7

HARAHWA, T. A., LAI YAU, T. H., LIM-COOKE, M.-S., AL-HADDI, S., ZEINAH, M. and HARKY, A. (2020). The optimal diagnostic methods for COVID-19. *Diagnosis* **7** 349–356. https://doi.org/10.1515/dx-2020-0058

HENSCHKE, P. J. (1993). Infections in the elderly. *Med. J. Aust.* **158** 830–834. https://doi.org/10.5694/j.1326-5377.1993.tb137672.x

HOLZMANN, H., MUNK, A. and GNEITING, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scand. J. Stat.* **33** 753–763. MR2300914 https://doi.org/10.1111/j.1467-9469.2006.00505.x

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

JIANG, X., COFFEE, M., BARI, A., WANG, J., JIANG, X., HUANG, J., SHI, J., DAI, J., CAI, J. et al. (2020). Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Comput. Mater. Continua* **62** 537–551. Available at arXiv:2020.010691. https://doi.org/10.32604/cmc

JOO, Y., BRUMBACK, B., LEE, K., YUN, S.-T., KIM, K.-H. and JOO, C. (2009). Clustering of temporal profiles using a Bayesian logistic mixture model: Analyzing groundwater level data to understand the characteristics of urban groundwater recharge. *J. Agric. Biol. Environ. Stat.* **14** 356–373. MR2750845 https://doi.org/10.1198/jabes.2009.07100

KUKAR, M., GUNČAR, G., VOVKO, T., PODNAR, S., ČERNELČ, P., BRVAR, M., ZALAZNIK, M., NOTAR, M., MOŠKON, S. et al. (2021). COVID-19 diagnosis by routine blood tests using machine learning. *Sci. Rep.* **11** 10738. https://doi.org/10.1038/s41598-021-90265-9

LU, Z. and SONG, X. (2012). Finite mixture varying coefficient models for analyzing longitudinal heterogenous data. *Stat. Med.* **31** 544–560. MR2892338 https://doi.org/10.1002/sim.4420

MA, P., HUANG, J. Z. and ZHANG, N. (2015). Efficient computation of smoothing splines via adaptive basis sampling. *Biometrika* **102** 631–645. MR3394280 https://doi.org/10.1093/biomet/asv009

MA, P. and ZHONG, W. (2008). Penalized clustering of large-scale functional data with multiple covariates. *J. Amer. Statist. Assoc.* **103** 625–636. MR2435467 https://doi.org/10.1198/016214508000000247

MA, X., GUO, W., GU, M., USVYAT, L., KOTANKO, P. and WANG, Y. (2024). Supplement to "A Nonparametric Mixed-Effects Mixture Model for Patterns of Clinical Measurements Associated with COVID-19." https://doi.org/10.1214/23-AOAS1871SUPPA, https://doi.org/10.1214/23-AOAS1871SUPPB, https://doi.org/10.1214/23-AOAS1871SUPPC

MACKOWIAK, P. A. (1997). *Fever: Basic Mechanisms and Management*, 2nd ed. Raven Press, New York.

MALIK, P., PATEL, U., MEHTA, D., PATEL, N., KELKAR, R., AKRMAH, M., GABRILOVE, J. L. and SACKS, H. (2021). Biomarkers and outcomes of COVID-19 hospitalisations: Systematic review and meta-analysis. *BMJ Evid.-Based Med.* **26** 107–108. https://doi.org/10.1136/bmjebm-2020-111536

MONAGHAN, C. K., LARKIN, J. W., CHAUDHURI, S., HAN, H., JIAO, Y., BERMUDEZ, K. M., WEINHANDL, E. D., DAHNE-STEUBER, I. A., BELMONTE, K. et al. (2021). Machine learning for prediction of patients on hemodialysis with an undetected SARS-CoV-2 infection. *Kidney*360 **2** 456–468. https://doi.org/10.34067/KID.0003802020

MUSGRAVE, T. and VERGHESE, A. (1990). Clinical features of pneumonia in the elderly. *Semin. Respir. Infect.* **5** 269–275.

NIDDK (2021). Kidney disease statistics for the united states NIDDK.

PIMENTEL, M. A. F., REDFERN, O. C., HATCH, R., YOUNG, J. D., TARASSENKO, L. and WATKINSON, P. J. (2020). Trajectories of vital signs in patients with COVID-19. *Resuscitation* **156** 99–106. https://doi.org/10.1016/j.resuscitation.2020.09.002

SIMON, B., RUBEY, H., TREIPL, A., GROMANN, M., HEMEDI, B., ZEHETMAYER, S. and KIRSCH, B. (2021). Haemodialysis patients show a highly diminished antibody response after COVID-19 mRNA vaccination compared with healthy controls. *Nephrol. Dial. Transplant.* **36** 1709–1716. https://doi.org/10.1093/ndt/gfab179

SOUZA, T. H., NADAL, J. A., NOGUEIRA, R. J. N., PEREIRA, R. M. and BRANDÃO, M. B. (2020). Clinical manifestations of children with COVID-19: A systematic review. *Pediatr. Pulmonol.* **55** 1892–1899. https://doi.org/10.1002/ppul.24885

SUN, X., ZHONG, W. and MA, P. (2021). An asymptotic and empirical smoothing parameters selection method for smoothing spline ANOVA models in large samples. *Biometrika* **108** 149–166. MR4226195 https://doi.org/10.1093/biomet/asaa047

TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Stat.* **34** 1265–1269. MR0155376 https://doi.org/10.1214/aoms/1177703862

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

USRDS (2020). Unites states renal data system annual data report.

WANG, S., YAO, W. and HUANG, M. (2014). A note on the identifiability of nonparametric and semiparametric mixtures of GLMs. *Statist. Probab. Lett.* **93** 41–45. MR3244553 https://doi.org/10.1016/j.spl.2014.06.010

WANG, Y. (1998a). Mixed effects smoothing spline analysis of variance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 159–174. MR1625640 https://doi.org/10.1111/1467-9868.00115

WANG, Y. (1998b). Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.* **93** 341–348. https://doi.org/10.1080/01621459.1998.10474115

WANG, Y. (2011). *Smoothing Splines: Methods and Applications. Monographs on Statistics and Applied Probability* **121**. CRC Press, Boca Raton, FL. MR2814838 https://doi.org/10.1201/b10954

WONG, K. Y., ZENG, D. and LIN, D. Y. (2022). Semiparametric latent-class models for multivariate longitudinal and survival data. *Ann. Statist.* **50** 487–510. MR4382025 https://doi.org/10.1214/21-aos2117

WU, J., ZHANG, P., ZHANG, L., MENG, W., LI, J., TONG, C., LI, Y., CAI, J., YANG, Z. et al. (2020). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. Preprint, Infectious Diseases (except HIV/AIDS). Available at https://doi.org/10.1101/2020.04.02.20051136.

XU, D. and WANG, Y. (2021). Low-rank approximation for smoothing spline via eigensystem truncation. *Stat* **10** Paper No. e355, 10. MR4235607 https://doi.org/10.1002/sta4.355

ZHU, C., BYRD, R. H., LU, P. and NOCEDAL, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Software* **23** 550–560. MR1671706 https://doi.org/10.1145/279232.279236

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2096–2122
https://doi.org/10.1214/24-AOAS1872

BENEFITS AND COSTS OF MATCHING PRIOR TO A DIFFERENCE IN DIFFERENCES ANALYSIS WHEN PARALLEL TRENDS DOES NOT HOLD

BY DAE WOONG HAM[1,a] AND LUKE MIRATRIX[2,b]

[1] *Department of Statistics, Harvard University,* [a]*daewoongham@g.harvard.edu*
[2] *Department of Education, Harvard Graduate School of Education,* [b]*lmiratrix@g.harvard.edu*

The consequence of a change in school leadership (e.g., principal turnover) on student achievement has important implications for education policy. The impact of such an event can be estimated via the popular difference in difference (DiD) estimator, where those schools with a turnover event are compared to a selected set of schools that did not have such an event. The strength of this comparison depends on the plausibility of the "parallel trends" assumption that the "treated group" of those schools which had leadership turnover, absent such turnover, would have changed "similarly" to those which did not. To bolster such a claim, one might generate a comparison group, via matching, that is similar to the treated group with respect to pretreatment outcomes and/or pretreatment covariates. Unfortunately, as has been previously pointed out, this intuitively appealing approach also has a cost in terms of bias. To assess the trade-offs of matching in our application, we first characterize the bias of matching prior to a DiD analysis under a linear structural model that allows for time-invariant observed and unobserved confounders with time-varying effects on the outcome. Given our framework, we verify that matching on baseline covariates generally reduces bias. We further show how additionally matching on pretreatment outcomes has both cost and benefit. First, matching on pretreatment outcomes partially balances unobserved confounders, which mitigates some bias. This reduction is proportional to the outcome's reliability, a measure of how coupled the outcomes are with the latent covariates. Offsetting these gains, matching also injects bias into the final estimate by undermining the second difference in the DiD via a regression-to-the-mean effect. Consequently, we provide heuristic guidelines for determining to what degree the bias reduction of matching is likely to outweigh the bias cost. We illustrate our guidelines by reanalyzing a principal turnover study that used matching prior to a DiD analysis and find that matching on both the pretreatment outcomes and observed covariates makes the estimated treatment effect more credible.

REFERENCES

ABADIE, A. (2005). Semiparametric difference-in-differences estimators. *Rev. Econ. Stud.* **72** 1–19. MR2116973 https://doi.org/10.1111/0034-6527.00321

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

ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267. MR2194325 https://doi.org/10.1111/j.1468-0262.2006.00655.x

ASHENFELTER, O. (1978). Estimating the effect of training programs on earnings. *Rev. Econ. Stat.* **60** 47–57.

BARNES, G. and BENJAMIN, C. (2007). The cost of teacher turnover in five school districts: A pilot study. National Commission on Teaching and America's Future.

BARTANEN, B., GRISSOM, J. A. and ROGERS, L. K. (2019). The impacts of principal turnover. *Educ. Eval. Policy Anal.* **41** 350–374. https://doi.org/10.3102/0162373719855044

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

Key words and phrases. Comparative interrupted time series, bias-bias tradeoff, linear structural equation model, latent confounder, reliability.

BOUTTELL, J., CRAIG, P., LEWSEY, J., ROBINSON, M. and POPHAM, F. (2018). Synthetic control methodology as a tool for evaluating population-level health interventions. *J. Epidemiol. Community Health* **72** 673–678. https://doi.org/10.1136/jech-2017-210106

BOYD, D., GROSSMAN, P., ING, M., LANKFORD, H., LOEB, S. and WYCKOFF, J. (2011). The influence of school administrators on teacher retention decisions. *Am. Educ. Res. J.* **48** 303–333. https://doi.org/10.3102/0002831210380788

CALLAWAY, B. and SANT'ANNA, P. H. C. (2021). Difference-in-differences with multiple time periods. *J. Econometrics* **225** 200–230. MR4328640 https://doi.org/10.1016/j.jeconom.2020.12.001

CARD, D. and SULLIVAN, D. (1987). Measuring the effect of subsidized training programs on movements in and out of employment. Working Paper No. 2173, National Bureau of Economic Research.

CHABÉ-FERRET, S. (2015). Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes. *J. Econometrics* **185** 110–123. MR3300339 https://doi.org/10.1016/j.jeconom.2014.09.013

CHABÉ-FERRET, S. (2017). Should we combine difference in differences with conditioning on pre-treatment outcomes?

COELLI, M. and GREEN, D. A. (2012). Leadership effects: School principals and student outcomes. *Econ. Educ. Rev.* **31** 92–109.

D'AMOUR, A., DING, P., FELLER, A., LEI, L. and SEKHON, J. (2021). Overlap in observational studies with high-dimensional covariates. *J. Econometrics* **221** 644–654. MR4215042 https://doi.org/10.1016/j.jeconom.2019.10.014

DAW, J. R. and HATFIELD, L. A. (2018). Matching and regression to the mean in difference-in-differences analysis. *Health Serv. Res.* **53** 4138–4156. https://doi.org/10.1111/1475-6773.12993

DING, P. and LI, F. (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Polit. Anal.* **27** 605–615. https://doi.org/10.1017/pan.2019.25

DING, P. and MIRATRIX, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. *J. Causal Inference* **3** 41–57. MR4289426 https://doi.org/10.1515/jci-2013-0021

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

GOLDRING, R. and TAIE, W. S. (2018). Principal attrition and mobility: Results from the 2016–17 principal follow-up survey. Technical Report, National Center for Education Statistics.

GRISSOM, J. A. and BARTANEN, B. (2019). Strategic retention: Principal effectiveness and teacher turnover in multiple-measure teacher evaluation systems. *Am. Educ. Res. J.* **56** 514–555. https://doi.org/10.3102/0002831218797931

GRISSOM, J. A., KALOGRIDES, D. and LOEB, S. (2015). Using student test scores to measure principal performance. *Educ. Eval. Policy Anal.* **37** 3–28. https://doi.org/10.3102/0162373714523831

GRISSOM, J. A., LOEB, S. and MASTER, B. (2013). Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. *Educ. Res.* **42** 433–444. https://doi.org/10.3102/0013189X13510020

HAM, D. W. and MIRATRIX, L. (2024a). Supplement A to "Benefits and costs of matching prior to a difference in difference analysis when parallel trends does not hold". https://doi.org/10.1214/24-AOAS1872SUPPA

HAM, D. W. and MIRATRIX, L. (2024b). Supplement B to "Benefits and costs of matching prior to a difference in difference analysis when parallel trends does not hold". https://doi.org/10.1214/24-AOAS1872SUPPB

HECKMAN, J., ICHIMURA, H., SMITH, J. and TODD, P. (1998). Characterizing selection bias using experimental data. *Econometrica* **66** 1017–1098. MR1639419 https://doi.org/10.2307/2999630

ILLENBERGER, N., SMALL, D. and SHAW, P. (2020). Impact of regression to the mean on the synthetic control method: Bias and sensitivity analysis. *Epidemiology* **31** 815–822. https://doi.org/10.1097/EDE.0000000000001252

IMAI, K., KIM, I. S. and WANG, E. H. (2023). Matching methods for causal inference with time-series cross-sectional data. *Amer. J. Polit. Sci.* **67** 587–605. https://doi.org/10.1111/ajps.12685

KIM, Y. and STEINER, P. M. (2021a). Causal graphical views of fixed effects and random effects models. *Br. J. Math. Stat. Psychol.* **74** 165–183. https://doi.org/10.1111/bmsp.12217

KIM, Y. and STEINER, P. M. (2021b). Gain scores revisited: A graphical models perspective. *Sociol. Methods Res.* **50** 1353–1375. MR4291995 https://doi.org/10.1177/0049124119826155

LENIS, D., EBNESAJJAD, C. F. and STUART, E. A. (2017). A doubly robust estimator for the average treatment effect in the context of a mean-reverting measurement error. *Biostatistics* **18** 325–337. MR3825123 https://doi.org/10.1093/biostatistics/kxw046

LINDNER, S. R. and MCCONNELL, K. J. (2018). Difference-in-differences and matching on outcomes: A tale of two unobservables. *Health Serv. Outcomes Res. Methodol.* 1–18.

PRASAD, K., VAIDYA, R. and VEMULA, A. (2016). An empirical analysis of the training program characteristics on training program effectiveness: A case study with reference to international agricultural research institute, Hyderabad. *J. Hum. Resour. Sustain. Stud.* **04** 143–154. https://doi.org/10.4236/jhrss.2016.43016

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

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

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. (2006). *Matched Sampling for Causal Effects*. Cambridge Univ. Press, Cambridge. MR2307965 https://doi.org/10.1017/CBO9780511810725

RUDOLPH, K. E. and STUART, E. A. (2018). Using sensitivity analyses for unobserved confounding to address covariate measurement error in propensity score methods. *Amer. J. Epidemiol.* **187** 604–613. https://doi.org/10.1093/aje/kwx248

RYAN, A., BURGESS, J. and DIMICK, J. (2015). Why we should not be indifferent to specification choices for difference-in-differences. *Health Serv. Res.* **50** 1211–1235. https://doi.org/10.1111/1475-6773.12270

SHPITSER, I., VANDERWEELE, T. and ROBINS, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*.

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

SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. MR1092986

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

TROCHIM, W. (2006). Types of reliability. Research Methods Knowledge Base, Web Center for Social Research Methods **04**.

TYLER, J. H., MURNANE, R. J. and WILLETT, J. B. (2000). Estimating the labor market signaling value of the GED*. *Q. J. Econ.* **115** 431–468. https://doi.org/10.1162/003355300554818

WEBB-VARGAS, Y., RUDOLPH, K. E., LENIS, D., MURAKAMI, P. and STUART, E. A. (2017). An imputation-based solution to using mismeasured covariates in propensity score analysis. *Stat. Methods Med. Res.* **26** 1824–1837. MR3687180 https://doi.org/10.1177/0962280215588771

WHARAM, J. F., LANDON, B. E., GALBRAITH, A. A., KLEINMAN, K. P., SOUMERAI, S. B. and ROSS-DEGNAN, D. (2007). Emergency department use and subsequent hospitalizations among members of a high-deductible health plan. *JAMA* **297** 1093–1102. https://doi.org/10.1001/jama.297.10.1093

WOOLDRIDGE, J. (2021). Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators.

ZELDOW, B. and HATFIELD, L. A. (2021). Confounding and regression adjustment in difference-in-differences studies. *Health Serv. Res.* **56** 932–941. https://doi.org/10.1111/1475-6773.13666

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2123–2146
https://doi.org/10.1214/24-AOAS1873
© Institute of Mathematical Statistics, 2024

A LATENT PROCESS MODEL FOR MONITORING PROGRESS TOWARD HARD-TO-MEASURE TARGETS WITH APPLICATIONS TO MENTAL HEALTH AND ONLINE EDUCATIONAL ASSESSMENTS

BY MINJEONG JEON[1,a] AND MICHAEL SCHWEINBERGER[2,b]

[1]*Department of Education, University of California, Los Angeles,* [a]*mjjeon@g.ucla.edu*
[2]*Department of Statistics, The Pennsylvania State University,* [b]*mus47@psu.edu*

The recent shift to remote learning and work has aggravated longstanding problems, such as the problem of monitoring the mental health of individuals and the progress of students toward learning targets. We introduce a novel latent process model with a view to monitoring the progress of individuals toward a hard-to-measure target of interest and measured by a set of variables. The latent process model is based on the idea of embedding both individuals and variables measuring progress toward the target of interest in a shared metric space, interpreted as an interaction map that captures interactions between individuals and variables. The fact that individuals are embedded in the same metric space as the target helps assess the progress of individuals toward the target. We demonstrate, with the help of simulations and applications, that the latent process model enables a novel look at mental health and online educational assessments in disadvantaged subpopulations.

REFERENCES

ALBERT, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *J. Educ. Stat.* **17** 251–269.

ANDERSEN, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika* **50** 3–16. MR0789214 https://doi.org/10.1007/BF02294143

BANG, H. J., LI, L. and FLYNN, K. (2022). Efficacy of an adaptive game-based math learning app to support personalized learning and improve early elementary school students' earning. *Early Child. Educ. J.* https://doi.org/10.1007/s10643-022-01332-3

BANSAK, C. and STARR, M. (2021). COVID-19 shocks to education supply: How 200,000 US households dealt with the sudden shift to distance learning. *Rev. Econ. Househ.* **19** 63–90.

BEEBER, L. S., SCHWARTZ, T. A., MARTINEZ, M. I., HOLDITCH-DAVIS, D., BLEDSOE, S. E., CANUSO, R. and LEWIS, V. S. (2014). Depressive symptoms and compromised parenting in low-income mothers of infants and toddlers: Distal and proximal risks. *Res. Nurs. Health* **37** 276–291. https://doi.org/10.1002/nur.21604

BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **9**. IMS, Hayward, CA. MR0882001

CAI, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika* **75** 581–612. MR2741489 https://doi.org/10.1007/s11336-010-9178-0

CURTIS, M. S. (2010). BUGS code for item response theory. *J. Stat. Softw.* **36** 1–34.

DALY, M., SUTIN, A. R. and ROBINSON, E. (2020). Longitudinal changes in mental health and the COVID-19 pandemic: Evidence from the UK Household Longitudinal Study. *Psychol. Med.* 1–10.

EFRON, B. (2023). *Exponential Families in Theory and Practice*. *Institute of Mathematical Statistics Textbooks* **16**. Cambridge Univ. Press, New York. MR4693002 https://doi.org/10.1017/9781108773157

EMBRETSON, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika* **56** 495–515.

ENGZELL, P., FREY, A. and VERHAGEN, M. D. (2021). Learning loss due to school closures during the COVID-19 pandemic. *Proc. Natl. Acad. Sci. USA* **118** e2022376118.

GELMAN, A. and HILL, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge Univ. Press, New York.

Key words and phrases. Latent space models, measurement models, item response models.

Handcock, M. S., Raftery, A. E. and Tantrum, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354. MR2364300 https://doi.org/10.1111/j.1467-985X.2007.00471.x

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

Holmes, E. A., O'Connor, R. C., Perry, V. H., Tracey, I., Wessely, S., Arseneault, L. and Everall, I. (2020). Multidisciplinary research priorities for the COVID-19 pandemic: A call for action for mental health science. *Lancet Psychiatry* **7** 547–560.

Huang, H.-Y. (2015). A multilevel higher order item response theory model for measuring latent growth in longitudinal data. *Appl. Psychol. Meas.* **39** 362–372. https://doi.org/10.1177/0146621614568112

Hunter, D. R., Krivitsky, P. N. and Schweinberger, M. (2012). Computational statistical methods for social network models. *J. Comput. Graph. Statist.* **21** 856–882. MR3005801 https://doi.org/10.1080/10618600.2012.732921

Jeon, M., Jin, I. H., Schweinberger, M. and Baugh, S. (2021). Mapping unobserved item-respondent interactions: A latent space item response model with interaction map. *Psychometrika* **86** 378–403. MR4291693 https://doi.org/10.1007/s11336-021-09762-5

Jeon, M. and Rabe-Hesketh, S. (2016). An autoregressive growth model for longitudinal item analysis. *Psychometrika* **81** 830–850. MR3535060 https://doi.org/10.1007/s11336-015-9489-2

Jeon, M. and Schweinberger, M. (2024). Supplement to "A latent process model for monitoring progress toward hard-to-measure targets with applications to mental health and online educational assessments." https://doi.org/10.1214/24-AOAS1873SUPP

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A. and Boguñá, M. (2010). Hyperbolic geometry of complex networks. *Phys. Rev. E* (3) **82** 036106, 18 pp. MR2787998 https://doi.org/10.1103/PhysRevE.82.036106

Krivitsky, P. N., Handcock, M. S., Raftery, A. E. and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Soc. Netw.* **31** 204–213. https://doi.org/10.1016/j.socnet.2009.04.001

Kuhfeld, M. et al. (2020). Projecting the potential impacts of COVID-19 school closures on academic achievement. *Educ. Res.* **49** 549–565.

Lubold, S., Chandrasekhar, A. G. and McCormick, T. H. (2023). Identifying the latent space geometry of network models through analysis of curvature. *J. Roy. Statist. Soc. Ser. B* **85** 240–292.

Pastor, D. A. and Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Appl. Psychol. Meas.* **30** 100–120. MR2225592 https://doi.org/10.1177/0146621605279761

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

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen, Denmark.

Santos, H. P. J., Kossakowski, J. J., Schwartz, T. A., Beeber, L. and Fried, E. I. (2018). Longitudinal network structure of depression symptoms and self-efficacy in low-income mothers. *PLoS ONE* **13** e0191675.

Schweinberger, M., Krivitsky, P. N., Butts, C. T. and Stewart, J. R. (2020). Exponential-family models of random graphs: Inference in finite, super and infinite population scenarios. *Statist. Sci.* **35** 627–662. MR4175389 https://doi.org/10.1214/19-STS743

Schweinberger, M. and Snijders, T. A. B. (2003). Settings in social networks: A measurement model. *Sociol. Method.* **33** 307–341.

Schweinberger, M. and Stewart, J. (2020). Concentration and consistency results for canonical and curved exponential-family models of random graphs. *Ann. Statist.* **48** 374–396. MR4065166 https://doi.org/10.1214/19-AOS1810

Segawa, E. (2005). A growth model for multilevel ordinal data. *J. Educ. Behav. Stat.* **30** 369–396.

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

Smith, A. L., Asta, D. M. and Calder, C. A. (2019). The geometry of continuous latent space models for network data. *Statist. Sci.* **34** 428–453. MR4017522 https://doi.org/10.1214/19-STS702

Sundberg, R. (2019). *Statistical Modelling by Exponential Families*. *Institute of Mathematical Statistics Textbooks* **12**. Cambridge Univ. Press, Cambridge. MR3969949 https://doi.org/10.1017/9781108604574

Vats, D. and Knudson, C. (2021). Revisiting the Gelman–Rubin diagnostic. *Statist. Sci.* **36** 518–529. MR4323050 https://doi.org/10.1214/20-sts812

Wang, C. and Nydick, S. W. (2020). On longitudinal item tesponse theory models: A didactic. *J. Educ. Behav. Stat.* **45** 339–368.

WATANABE, S. (2013). A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14** 867–897. MR3049492

WILSON, M., ZHENG, X. and MCGUIRE, L. W. (2012). Formulating latent growth using an explanatory item response model approach. *J. Appl. Meas.* **13** 1–22.

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2147–2165
https://doi.org/10.1214/24-AOAS1874
© Institute of Mathematical Statistics, 2024

EXPOSURE EFFECTS ON COUNT OUTCOMES WITH OBSERVATIONAL DATA, WITH APPLICATION TO INCARCERATED WOMEN

BY BONNIE E. SHOOK-SA[1,a], MICHAEL G. HUDGENS[1,b], ANDREA K. KNITTEL[2,c], ANDREW EDMONDS[3,f], CATALINA RAMIREZ[2,d], STEPHEN R. COLE[3,g], MARDGE COHEN[4,h], ADEBOLA ADEDIMEJI[5,i], TONYA TAYLOR[6,j], KATHERINE G. MICHEL[7,k], ANDREA KOVACS[8,l], JENNIFER COHEN[9,m], JESSICA DONOHUE[10,n], ANTONINA FOSTER[11,o], MARGARET A. FISCHL[12,p], DUSTIN LONG[13,q] AND ADAORA A. ADIMORA[2,e]

[1]*Department of Biostatistics, University of North Carolina at Chapel Hill,* [a]*bshooksa@email.unc.edu,* [b]*mhudgens@email.unc.edu*

[2]*School of Medicine, University of North Carolina at Chapel Hill,* [c]*andrea_knittel@med.unc.edu,* [d]*catalina_ramirez@med.unc.edu,* [e]*adimora@med.unc.edu*

[3]*Department of Epidemiology, University of North Carolina at Chapel Hill,* [f]*aedmonds@email.unc.edu,* [g]*cole@unc.edu*

[4]*Stroger Hospital,* [h]*mardge.cohen@gmail.com*

[5]*Albert Einstein College of Medicine,* [i]*adebola.adedimeji@einsteinmed.org*

[6]*SUNY Downstate Medical Center,* [j]*Tonya.Taylor@downstate.edu*

[7]*Department of Infectious Diseases, Georgetown University,* [k]*kgm52@georgetown.edu*

[8]*Keck School of Medicine, University of Southern California,* [l]*akovacs@usc.edu*

[9]*Department of Medicine, University of California, San Francisco,* [m]*jennifer.cohen@ucsf.edu*

[10]*Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health,* [n]*jdonohu7@jhu.edu*

[11]*Department of Medicine, Emory University,* [o]*antonina.g.jackson@emory.edu*

[12]*Division of Infectious Diseases, University of Miami Miller School Medicine,* [p]*mfischl@med.miami.edu*

[13]*The University of Alabama at Birmingham,* [q]*dmlong@uab.edu*

Causal inference methods can be applied to estimate the effect of a point exposure or treatment on an outcome of interest using data from observational studies. For example, in the Women's Interagency HIV Study, it is of interest to understand the effects of incarceration on the number of sexual partners and the number of cigarettes smoked after incarceration. In settings like this where the outcome is a count, the estimand is often the causal mean ratio, that is, the ratio of the counterfactual mean count under exposure to the counterfactual mean count under no exposure. This paper considers estimators of the causal mean ratio based on inverse probability of treatment weights, the parametric g-formula, and doubly robust estimation, each of which can account for overdispersion, zero-inflation, and heaping in the measured outcome. Methods are compared in simulations and are applied to data from the Women's Interagency HIV Study.

REFERENCES

ADIMORA, A. A., RAMIREZ, C., BENNING, L., GREENBLATT, R. M., KEMPF, M.-C., TIEN, P. C., KASSAYE, S. G., ANASTOS, K., COHEN, M. et al. (2018). Cohort profile: The Women's Interagency HIV Study (WIHS). *Int. J. Epidemiol.* **47** 393–394.

ALBERT, J. M., WANG, W. and NELSON, S. (2014). Estimating overall exposure effects for zero-inflated regression models with application to dental caries. *Stat. Methods Med. Res.* **23** 257–278. MR3215053 https://doi.org/10.1177/0962280211407800

BAILEY, Z. D., OKECHUKWU, C., KAWACHI, I. and WILLIAMS, D. R. (2015). Incarceration and current tobacco smoking among black and Caribbean black Americans in the national survey of American life. *Amer. J. Publ. Health* **105** 2275–2282. https://doi.org/10.2105/AJPH.2015.302772

Key words and phrases. Data heaping, doubly robust estimation, inverse probability weighting, overdispersion, parametric g-formula, zero-inflation.

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–973. MR2216189 https://doi.org/10.1111/j.1541-0420.2005.00377.x

BANGSBERG, D. R., HECHT, F. M., CHARLEBOIS, E. D., CHESNEY, M. and MOSS, A. (2001). Comparing objective measures of adherence to HIV antiretroviral therapy: Electronic medication monitors and unannounced pill counts. *AIDS Behav.* **5** 275–281.

BENECHA, H. K., NEELON, B., DIVARIS, K. and PREISSER, J. S. (2017). Marginalized mixture models for count data from multiple source populations. *J. Stat. Distrib. Appl.* **4** 1–17.

BINSWANGER, I. A., CARSON, E. A., KRUEGER, P. M., MUELLER, S. R., STEINER, J. F. and SABOL, W. J. (2014). Prison tobacco control policies and deaths from smoking in United States prisons: Population based retrospective analysis. *BMJ* **349** 1–12. https://doi.org/10.1136/bmj.g4542

BODNAR, L. M., DAVIDIAN, M., SIEGA-RIZ, A. M. and TSIATIS, A. A. (2004). Marginal structural models for analyzing causal effects of time-dependent treatments: An application in perinatal epidemiology. *Amer. J. Epidemiol.* **159** 926–934. https://doi.org/10.1093/aje/kwh131

BÖHNING, D., DIETZ, E., SCHLATTMANN, P., MENDONCA, L. and KIRCHNER, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *J. Roy. Statist. Soc. Ser. A* **162** 195–209.

COLE, S. R. and HERNÁN, M. A. (2008). Constructing inverse probability weights for marginal structural models. *Amer. J. Epidemiol.* **168** 656–664. https://doi.org/10.1093/aje/kwn164

CROPSEY, K., ELDRIDGE, G., WEAVER, M., VILLALOBOS, G., STITZER, M. and BEST, A. (2008). Smoking cessation intervention for female prisoners: Addressing an urgent public health need. *Amer. J. Publ. Health* **98** 1894–1901. https://doi.org/10.2105/AJPH.2007.128207

FUNK, M. J., WESTREICH, D., WIESEN, C., STÜRMER, T., BROOKHART, M. A. and DAVIDIAN, M. (2011). Doubly robust estimation of causal effects. *Amer. J. Epidemiol.* **173** 761–767. https://doi.org/10.1093/aje/kwq439

GARCIA-AYMERICH, J., VARRASO, R., DANAEI, G., CAMARGO, J. C. A. and HERNÁN, M. A. (2013). Incidence of adult-onset asthma after hypothetical interventions on body mass index and physical activity: An application of the parametric G-formula. *Amer. J. Epidemiol.* **179** 20–26.

HARAWA, N. and ADIMORA, A. (2008). Incarceration African Americans and HIV: Advancing a research agenda. *J. Natl. Med. Assoc.* **100** 57–63.

HERNÁN, M. A., BRUMBACK, B. and ROBINS, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* **11** 561–570. https://doi.org/10.1097/00001648-200009000-00012

HERNÁN, M. Á. and ROBINS, J. M. (2020). *Causal Inference*: *What If*. CRC Press/CRC, Boca Raton, FL.

HIV.GOV (2020). How does smoking affect people with HIV? Available at: https://www.hiv.gov/hiv-basics/staying-in-hiv-care/other-related-health-issues/smoking.

KAJSTURA, A. (2019). In women's mass incarceration: The whole pie 2019. Prison policy initiative, Northampton, MA.

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458 https://doi.org/10.1214/07-STS227

KAUFFMAN, R. M., FERKETICH, A. K. and WEWERS, M. E. (2008). Tobacco policy in American prisons, 2007. *Tob. Control* **17** 357–360. https://doi.org/10.1136/tc.2007.024448

KENNEDY, S. M., DAVIS, S. P. and THORNE, S. L. (2015). Smoke-free policies in U.S. prisons and jails: A review of the literature. *Nicotine Tob. Res.* **17** 629–635. https://doi.org/10.1093/ntr/ntu225

KLESGES, R. C., DEBON, M. and RAY, J. W. (1995). Are self-reports of smoking rate biased? Evidence from the second national health and nutrition examination survey. *J. Clin. Epidemiol.* **48** 1225–1233. https://doi.org/10.1016/0895-4356(95)00020-5

KNITTEL, A. K., SHOOK-SA, B. E., RUDOLPH, J., EDMONDS, A., RAMIREZ, C., COHEN, M., ADEDIMEJI, A., TAYLOR, T., MICHEL, K. G. et al. (2020). Incarceration and number of sexual partners after incarceration among vulnerable US women, 2007–2017. *Amer. J. Publ. Health* **110** S100–S108. https://doi.org/10.2105/AJPH.2019.305410

LONG, D. L., PREISSER, J. S., HERRING, A. H. and GOLIN, C. E. (2014). A marginalized zero-inflated Poisson regression model with overall exposure effects. *Stat. Med.* **33** 5151–5165. MR3276526 https://doi.org/10.1002/sim.6293

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. https://doi.org/10.1002/sim.1903

MDODO, R., FRAZIER, E. L., DUBE, S. R., MATTSON, C. L., SUTTON, M. Y., BROOKS, J. T. and SKARBINSKI, J. (2015). Cigarette smoking prevalence among adults with HIV compared with the general adult population in the United States: Cross-sectional surveys. *Ann. Intern. Med.* **162** 335–344. https://doi.org/10.7326/M14-0954

MORRIS, T. P., WHITE, I. R. and CROWTHER, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* **38** 2074–2102. MR3937487 https://doi.org/10.1002/sim.8086

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

NORTON, E. C., DOWD, B. E. and MACIEJEWSKI, M. L. (2018). Odds ratios-current best practice and use. *JAMA* **320** 84–85. https://doi.org/10.1001/jama.2018.6971

PREISSER, J. S., DAS, K., BENECHA, H. and STAMM, J. W. (2016a). Logistic regression for dichotomized counts. *Stat. Methods Med. Res.* **25** 3038–3056. MR3572897 https://doi.org/10.1177/0962280214536893

PREISSER, J. S., DAS, K., LONG, D. L. and DIVARIS, K. (2016b). Marginalized zero-inflated negative binomial regression with application to dental caries. *Stat. Med.* **35** 1722–1735. MR3513480 https://doi.org/10.1002/sim.6804

ROBERTS, J. M. JR. and BREWER, D. D. (2001). Measures and tests of heaping in discrete quantitative distributions. *J. Appl. Stat.* **28** 887–896. MR1863441 https://doi.org/10.1080/02664760120074960

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. (1998). Marginal structural models. In 1997 *Proceedings of the American Statistical Association, Section on Bayesian Statistical Science* 1–10.

ROBINS, J. M. HERNÁN, M. Á. 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. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

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

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Classics Library*. Wiley-Interscience, New York. Reprint of the 1987 edition [John Wiley & Sons, Inc., New York; MR899519]. MR2117498

SAUL, B. and HUDGENS, M. (2020). The calculus of M-estimation in R with geex. *J. Stat. Softw.* **92** 1–15.

SCHNITZER, M. E., VAN DER LAAN, M. J., MOODIE, E. E. M. and PLATT, R. W. (2014). Effect of breastfeeding on gastrointestinal infection in infants: A targeted maximum likelihood approach for clustered longitudinal data. *Ann. Appl. Stat.* **8** 703–725. MR3262531 https://doi.org/10.1214/14-AOAS727

SHOOK-SA, B. E., HUDGENS, M. G., KNITTEL, A. K., EDMONDS, A., RAMIREZ, C., COLE, S. R., COHEN, M., ADEDIMEJI, A., TAYLOR, T. et al. (2024). Supplement to "Exposure effects on count outcomes with observational data, with application to incarcerated women." https://doi.org/10.1214/24-AOAS1874SUPP

SHU, D. and YI, G. Y. (2019). Causal inference with measurement error in outcomes: Bias analysis and estimation methods. *Stat. Methods Med. Res.* **28** 2049–2068. MR3977092 https://doi.org/10.1177/0962280217743777

SINGH, K. K., SUCHINDRAN, C. M. and SINGH, R. S. (1994). Smoothed breastfeeding durations and waiting time to conception. *Soc. Biol.* **41** 229–239. https://doi.org/10.1080/19485565.1994.9988874

SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. MR1092986

STEFANSKI, L. A. and BOOS, D. D. (2002). The calculus of M-estimation. *Amer. Statist.* **56** 29–38. MR1939394 https://doi.org/10.1198/000313002753631330

TAUBMAN, S. L., ROBINS, J. M., MITTLEMAN, M. A. and HERNÁN, M. A. (2009). Intervening on risk factors for coronary heart disease: An application of the parametric g-formula. *Int. J. Epidemiol.* **38** 1599–1611. https://doi.org/10.1093/ije/dyp192

VANDERWEELE, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology* **20** 880–883. https://doi.org/10.1097/EDE.0b013e3181bd5638

WAERNBAUM, I. (2012). Model misspecification and robustness in causal inference: Comparing matching with doubly robust estimation. *Stat. Med.* **31** 1572–1581. MR2947528 https://doi.org/10.1002/sim.4496

WANG, H. and HEITJAN, D. F. (2008). Modeling heaping in self-reported cigarette counts. *Stat. Med.* **27** 3789–3804. MR2526609 https://doi.org/10.1002/sim.3281

WIEDERMAN, M. W. (1997). The truth must be in here somewhere: Examining the gender discrepancy in self-reported lifetime number of sex partners. *J. Sex Res.* **34** 375–386.

YOUNG, J. G., CAIN, L. E., ROBINS, J. M., O'REILLY, E. J. and HERNÁN, M. A. (2011). Comparative effectiveness of dynamic treatment regimes: An application of the parametric g-formula. *Stat. Biosci.* **3** 119–143.

ZHANG, J. (2018). Prison smoking bans in the United States: Current policy, impact and obstacle. *J. Hospital Manag. Health Policy* **2** 1–4.

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2166–2186
https://doi.org/10.1214/24-AOAS1875

SUPPORT VECTOR MACHINE FOR DYNAMIC SURVIVAL PREDICTION WITH TIME-DEPENDENT COVARIATES

BY WENYI XIE[1,a], DONGLIN ZENG[2,b] AND YUANJIA WANG[3,c]

[1] *Department of Biostatistics, University of North Carolina at Chapel Hill* , [a] *xiewenyi@live.unc.edu*
[2] *Department of Biostatistics, University of Michigan* , [b] *dzeng@umich.edu*
[3] *Department of Biostatistics, Mailman School of Public Health, Columbia University,* [c] *yw2016@cumc.columbia.edu*

Predicting time-to-event outcomes using time-dependent covariates is a challenging problem. Many machine learning approaches, such as tree-based methods and support vector regression, predominantly utilize only baseline covariates. Only a few methods can incorporate time-dependent covariates, but they often lack theoretical justification. In this paper we present a new framework for event time prediction, leveraging the support vector machines to forecast the associated counting processes. Utilizing the kernel trick, we accommodate nonlinear functions in both time and covariate spaces. Subsequently, we use a chain algorithm to predict future events. Theoretical analysis proves that our method is equivalent to comparing time-varying hazard rates among at-risk subjects, and we obtain the convergence rate of the resulting prediction loss. Through simulation studies and a case study on Huntington's disease, we demonstrate the superior performance of our approach compared to alternative methods based on machine learning, deep learning, and statistical models.

REFERENCES

AGUS, F., CRESPO, D., MYERS, R. H. and LABADORF, A. (2019). The caudate nucleus undergoes dramatic and unique transcriptional changes in human prodromal Huntington's disease brain. *BMC Med. Genom.* **12** 1–17.

BACCHETTI, P. and SEGAL, M. R. (1995). Survival trees with time-dependent covariates: Application to estimating changes in the incubation period of AIDS. *Lifetime Data Anal.* **1** 35–47. https://doi.org/10.1007/BF00985256

BOU-HAMAD, I., LAROCQUE, D. and BEN-AMEUR, H. (2011a). Discrete-time survival trees and forests with time-varying covariates: Application to bankruptcy data. *Stat. Model.* **11** 429–446. MR2907837 https://doi.org/10.1177/1471082X1001100503

BOU-HAMAD, I., LAROCQUE, D. and BEN-AMEUR, H. (2011b). A review of survival trees. *Stat. Surv.* **5** 44–71. MR3018509 https://doi.org/10.1214/09-SS047

CHEN, G. H. (2020). Deep kernel survival analysis and subject-specific survival time prediction intervals. In *Machine Learning for Healthcare Conference* 537–565. PMLR.

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758 https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

FARAGGI, D. and SIMON, R. (1995). A neural network model for survival data. *Stat. Med.* **14** 73–82. https://doi.org/10.1002/sim.4780140108

FU, W. and SIMONOFF, J. S. (2017). Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics* **18** 352–369. MR3825124 https://doi.org/10.1093/biostatistics/kxw047

GENSHEIMER, M. F. and NARASIMHAN, B. (2019). A scalable discrete-time survival model for neural networks. *PeerJ* **7** e6257. https://doi.org/10.7717/peerj.6257

GOLDBERG, Y. and KOSOROK, M. R. (2017). Support vector regression for right censored data. *Electron. J. Stat.* **11** 532–569. MR3619316 https://doi.org/10.1214/17-EJS1231

HOTHORN, T., BÜHLMANN, P., DUDOIT, S., MOLINARO, A. and VAN DER LAAN, M. J. (2006). Survival ensembles. *Biostatistics* **7** 355–373. https://doi.org/10.1093/biostatistics/kxj011

HOTHORN, T., HORNIK, K. and ZEILEIS, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Statist.* **15** 651–674. MR2291267 https://doi.org/10.1198/106186006X133933

Key words and phrases. Cox model, survival analysis, dynamic prediction, Huntington's Disease, machine learning, reproducing kernel Hilbert space.

HOTHORN, T., LAUSEN, B., BENNER, A. and RADESPIEL-TRÖGER, M. (2004). Bagging survival trees. *Stat. Med.* **23** 77–91. https://doi.org/10.1002/sim.1593

HUANG, X., CHEN, S. and SOONG, S. J. (1998). Piecewise exponential survival trees with time-dependent covariates. *Biometrics* **54** 1420–1433.

ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* **2** 841–860. MR2516796 https://doi.org/10.1214/08-AOAS169

JING, B., ZHANG, T., WANG, Z., JIN, Y., LIU, K., QIU, W., KE, L., SUN, Y., HE, C. et al. (2019). A deep survival analysis method based on ranking. *Artif. Intell. Med.* **98** 1–9.

KALBFLEISCH, J. D. and PRENTICE, R. L. (2011). *The Statistical Analysis of Failure Time Data. Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0570114

KATZMAN, J., SHAHAM, U., CLONINGER, A., BATES, J., JIANG, T. and KLUGER, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**.

KHAN, F. M. and ZUBEK, V. B. (2008). Support vector regression for censored data (SVRc): A novel tool for survival analysis. In 2008 *Eighth IEEE International Conference on Data Mining* 863–868. https://doi.org/10.1109/ICDM.2008.50

KVAMME, H. and BORGAN, Ø. (2021). Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Anal.* **27** 710–736. MR4330436 https://doi.org/10.1007/s10985-021-09532-6

KVAMME, H., BORGAN, Ø. and SCHEEL, I. (2019). Time-to-event prediction with neural networks and Cox regression. *J. Mach. Learn. Res.* **20** Paper No. 129, 30 pp. MR4002883

LEE, C., YOON, J. and SCHAAR, M. V. D. (2020). Dynamic-DeepHit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Trans. Biomed. Eng.* **67** 122–133. https://doi.org/10.1109/TBME.2019.2909027

LEE, C., ZAME, W., YOON, J. and VAN DER SCHAAR, M. (2018). DeepHit: A deep learning approach to survival analysis with competing risks. *Proc. AAAI Conf. Artif. Intell.* **32**. https://doi.org/10.1609/aaai.v32i1.11842

LI, F., LI, K., LI, C., LUO, S. et al. (2019). Predicting the risk of Huntington's disease with multiple longitudinal biomarkers. *J. Huntington's Dis.* **8** 323–332.

PAULSEN, J. S., LANGBEHN, D. R., STOUT, J. C., AYLWARD, E., ROSS, C. A., NANCE, M., GUTTMAN, M., JOHNSON, S., MACDONALD, M. et al. (2008). Detection of Huntington's disease decades before diagnosis: The predict-HD study. *J. Neurol. Neurosurg. Psychiatry* **79** 874–880.

RIPLEY, R. M., HARRIS, A. L. and TARASSENKO, L. (2004). Non-linear survival analysis using neural networks. *Stat. Med.* **23** 825–842. https://doi.org/10.1002/sim.1655

ROBINS, J. and TSIATIS, A. A. (1992). Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika* **79** 311–319. MR1185133 https://doi.org/10.1093/biomet/79.2.311

SANZ, H., REVERTER, F. and VALIM, C. (2020). Enhancing SVM for survival data using local invariances and weighting. *BMC Bioinform.* **21** 193. https://doi.org/10.1186/s12859-020-3481-2

SEGAL, M. R. (1988). Regression trees for censored data. *Biometrics* **44** 35–47.

SHIVASWAMY, P. K., CHU, W. and JANSCHE, M. (2007). A support vector approach to censored targets. In *Seventh IEEE International Conference on Data Mining* (*ICDM* 2007) 655–660. IEEE, New York. https://doi.org/10.1109/ICDM.2007.93

SUN, Y., CHIOU, S. H. and WANG, M.-C. (2020). ROC-guided survival trees and ensembles. *Biometrics* **76** 1177–1189. MR4186834 https://doi.org/10.1111/biom.13213

VAN BELLE, V., PELCKMANS, K., HUFFEL, S. V. and SUYKENS, J. A. K. (2011). Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artif. Intell. Med.* **53** 107–118. https://doi.org/10.1016/j.artmed.2011.06.006

WALLACE, M. L. (2014). Time-dependent tree-structured survival analysis with unbiased variable selection through permutation tests. *Stat. Med.* **33** 4790–4804. MR3274512 https://doi.org/10.1002/sim.6261

WANG, Y., CHEN, T. and ZENG, D. (2016). Support vector hazards machine: A counting process framework for learning risk scores for censored outcomes. *J. Mach. Learn. Res.* **17** Paper No. 167, 37 pp. MR3555058 https://doi.org/10.1016/j.insmatheco.2015.11.005

WEI, L.-J. (1992). The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Stat. Med.* **11** 1871–1879.

XIE, W., ZENG, D. and WANG, Y. (2024). Supplement to "Support vector machine for dynamic survival prediction with time-dependent covariates." https://doi.org/10.1214/24-AOAS1875SUPPA, https://doi.org/10.1214/24-AOAS1875SUPPB

ZENG, D. and LIN, D. Y. (2007). Efficient estimation for the accelerated failure time model. *J. Amer. Statist. Assoc.* **102** 1387–1396. MR2412556 https://doi.org/10.1198/016214507000001085

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2187–2206
https://doi.org/10.1214/24-AOAS1876
© Institute of Mathematical Statistics, 2024

NONCONVEX SVM FOR CANCER DIAGNOSIS BASED ON MORPHOLOGIC FEATURES OF TUMOR MICROENVIRONMENT

BY SEAN KENT[1,a] AND MENGGANG YU[2,b]

[1]*Department of Statistics, University of Wisconsin–Madison,* [a]*skent259@gmail.com*
[2]*Department of Biostatistics, University of Michigan,* [b]*menggang@umich.edu*

The surroundings of a cancerous tumor impact how it grows and develops in humans. New data from early breast cancer patients contains information on the collagen fibers surrounding the tumorous tissue—offering hope of finding additional biomarkers for diagnosis and prognosis—but poses two challenges for typical analysis. Each image section contains information on hundreds of fibers, and each tissue has multiple image sections contributing to a single prediction of tumor vs. nontumor. This nested relationship of fibers within image spots within tissue samples requires a specialized analysis approach.

We devise a novel support vector machine (SVM)-based predictive algorithm for this data structure. By treating the collection of fibers as a probability distribution, we can measure similarities between the collections through a flexible kernel approach. By assuming the relationship of tumor status between image sections and tissue samples, the constructed SVM problem is nonconvex, and traditional algorithms can not be applied. We propose two algorithms that exchange computational accuracy and efficiency to manage data of all sizes. The predictive performance of both algorithms is evaluated on the collagen fiber data set and additional simulation scenarios. We offer reproducible implementations of both algorithms of this approach in the R package `mildsvm`.

REFERENCES

ALPAYDIN, E., CHEPLYGINA, V., LOOG, M. and TAX, D. M. J. (2015). Single- vs. multiple-instance classification. *Pattern Recognit.* **48** 2831–2838. https://doi.org/10.1016/j.patcog.2015.04.006

AMINOLOLAMA-SHAKERI, S., FLOWERS, C. I., MCLAREN, C. E., WISNER, D. J., DE GUZMAN, J., CAMPBELL, J. E., BASSETT, L. W., OJEDA-FOURNIER, H., GERLACH, K. et al. (2017). Can radiologists predict the presence of ductal carcinoma in situ and invasive breast cancer? *Amer. J. Roentgenol.* **208** 933–939. https://doi.org/10.2214/AJR.16.16073

ANDERSON, N. M. and SIMON, M. C. (2020). The tumor microenvironment. *Curr. Biol.* **30** R921–R925. https://doi.org/10.1016/j.cub.2020.06.081

ANDREWS, S., TSOCHANTARIDIS, I. and HOFMANN, T. (2003). Support vector machines for multiple-instance learning. *Adv. Neural Inf. Process. Syst.* **15** 577–584.

ARENDT, L. M., RUDNICK, J. A., KELLER, P. J. and KUPERWASSER, C. (2010). Stroma in breast development and disease. *Semin. Cell Dev. Biol.* **21** 11–18. https://doi.org/10.1016/j.semcdb.2009.10.003

BAGHBAN, R., ROSHANGAR, L., JAHANBAN-ESFAHLAN, R., SEIDI, K., EBRAHIMI-KALAN, A., JAYMAND, M., KOLAHIAN, S., JAVAHERI, T. and ZARE, P. (2020). Tumor microenvironment complexity and therapeutic implications at a glance. *Cell Commun. Signal.* **18**. https://doi.org/10.1186/s12964-020-0530-4

BEJARANO, L., JORDÃO, M. J. C. and JOYCE, J. A. (2021). Therapeutic targeting of the tumor microenvironment. *Cancer Discov.* **11** 933–959. https://doi.org/10.1158/2159-8290.CD-20-1808

BOSER, B. E., GUYON, I. M. and VAPNIK, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* 144–152. https://doi.org/10.1145/130385.130401

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

BURGES, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **2** 121–167. https://doi.org/10.1023/A:1009715923555

Key words and phrases. Breast cancer, support vector machines, weakly supervised, functional data.

CAMPANELLA, G., HANNA, M. G., GENESLAW, L., MIRAFLOR, A., WERNECK KRAUSS SILVA, V., BUSAM, K. J., BROGI, E., REUTER, V. E., KLIMSTRA, D. S. et al. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25** 1301–1309. https://doi.org/10.1038/s41591-019-0508-1

CHATALIC, A., SCHREUDER, N., ROSASCO, L. and RUDI, A. (2022). Nyström kernel mean embeddings. In *International Conference on Machine Learning* 3006–3024. PMLR, Baltimore, MD, USA.

CHEN, J., ZENG, H., ZHANG, C., SHI, Z., DEKKER, A., WEE, L. and BERMEJO, I. (2022). Lung cancer diagnosis using deep attention based multiple instance learning and radiomics. *Med. Phys.* **49** 3134–3143. https://doi.org/10.1002/mp.15539

CHEN, P.-Y., CHEN, C.-C., YANG, C.-H., CHANG, S.-M. and LEE, K.-J. (2017). milr: Multiple-instance logistic regression with lasso penalty. *R J.* **9** 446. https://doi.org/10.32614/RJ-2017-013

CHEN, X., NADIARYNKH, O., PLOTNIKOV, S. and CAMPAGNOLA, P. J. (2012). Second harmonic generation microscopy for quantitative analysis of collagen fibrillar structure. *Nat. Protoc.* **7** 654–669. https://doi.org/10.1038/nprot.2012.009

CONKLIN, M. W., GANGNON, R. E., SPRAGUE, B. L., GEMERT, L. V., HAMPTON, J. M., ELICEIRI, K. W., BREDFELDT, J. S., LIU, Y., SURACHAICHARN, N. et al. (2018). Collagen alignment as a predictor of recurrence after ductal carcinoma in situ. *Cancer Epidemiol. Biomark. Prev.* **27** 138–145. https://doi.org/10.1158/1055-9965.EPI-17-0720

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. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

DIETTERICH, T. G., LATHROP, R. H. and LOZANO-PÉREZ, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence* **89** 31–71. https://doi.org/10.1016/S0004-3702(96)00034-3

ERTEKIN, S., BOTTOU, L. and GILES, C. L. (2010). Nonconvex online support vector machines. *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 368–381. MR2841444 https://doi.org/10.1007/978-1-84996-098-4

GUAN, L., SUN, T., QIAO, L.-B., YANG, Z.-H., LI, D.-S., GE, K.-S. and LU, X.-C. (2020). An efficient parallel and distributed solution to nonconvex penalized linear SVMs. *Front. Inf. Technol. & Electron. Eng.* **21** 587–603.

GUROBI OPTIMIZATION L. (2021). Mixed-integer programming (MIP)—a primer on the basics.

KENT, S. and YU, M. (2024). Supplement to “Nonconvex SVM for cancer diagnosis based on morphologic features of tumor microenvironment.” https://doi.org/10.1214/24-AOAS1876SUPPA, https://doi.org/10.1214/24-AOAS1876SUPPB

KIM, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comput. Statist. Data Anal.* **53** 3735–3745. MR2749918 https://doi.org/10.1016/j.csda.2009.04.009

LANGE, K. (2016). *MM Optimization Algorithms*. SIAM, Philadelphia, PA. MR3522165 https://doi.org/10.1137/1.9781611974409.ch1

LAPORTE, L., FLAMARY, R., CANU, S., DÉJEAN, S. and MOTHE, J. (2013). Nonconvex regularizations for feature selection in ranking with sparse SVM. *IEEE Trans. Neural Netw. Learn. Syst.* **25** 1118–1130. Publisher: IEEE.

LAZIMY, R. (1982). Mixed-integer quadratic programming. *Math. Program.* **22** 332–349. MR0646573 https://doi.org/10.1007/BF01581047

LI, Y.-F., TSANG, I. W., KWOK, J. T. and ZHOU, Z.-H. (2013). Convex and scalable weakly labeled SVMs. *J. Mach. Learn. Res.* **14** 2151–2188. MR3104505

LIN, H.-T., LEE, S., BUI, N. and HONAVAR, V. (2013). Learning classifiers from distributional data. In 2013 *IEEE International Congress on Big Data*. 302–309. MR3310603 https://doi.org/10.1109/BigData.Congress.2013.47

LIU, H., YAO, T. and LI, R. (2016). Global solutions to folded concave penalized nonconvex learning. *Ann. Statist.* **44** 629–659. MR3476612 https://doi.org/10.1214/15-AOS1380

MINH, H. Q., NIYOGI, P. and YAO, Y. (2006). Mercer’s theorem, feature maps, and smoothing. In *Learning Theory*. *Lecture Notes in Computer Science* **4005** 154–168. Springer, Berlin. MR2280604 https://doi.org/10.1007/11776420_14

MITCHELL, M. (1998). *An Introduction to Genetic Algorithms*. MIT press, Cambridge.

MUANDET, K., FUKUMIZU, K., DINUZZO, F. and SCHÖLKOPF, B. (2012). Learning from distributions via support measure machines. *Adv. Neural Inf. Process. Syst.* **25** 10–18.

MUANDET, K., FUKUMIZU, K., SRIPERUMBUDUR, B. and SCHÖLKOPF, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Found. Trends Mach. Learn.* **10** 1–141. https://doi.org/10.1561/2200000060

PIA, A. D., DEY, S. S. and MOLINARO, M. (2017). Mixed-integer quadratic programming is in NP. *Math. Program.* **162** 225–240. MR3612939 https://doi.org/10.1007/s10107-016-1036-0

PLATT, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report No. MSR-TR-98-14, Microsoft.

POLI, R., KENNEDY, J. and BLACKWELL, T. (2007). Particle swarm optimization: An overview. *Swarm Intell.* **1** 33–57. https://doi.org/10.1007/s11721-007-0002-0

POWERS, D. M. (2011). Evaluation: From predcision, recall and F-factor to ROC, informedness, markedness & correlation. *Mach. Learn. Technol.* **2** 37–63.

RAHIMI, A. and RECHT, B. (2008). Random features for large-scale kernel machines. *Adv. Neural Inf. Process. Syst.* **20** 1177–1184.

RAMSAY, J. O. (2006). Functional data analysis. *Encycl. Statist. Sci.* https://doi.org/10.1002/0471667196.ess3138

RAY, S. and CRAVEN, M. (2005). Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of the* 22*nd International Conference on Machine Learning* 697–704. https://doi.org/10.1145/1102351.1102439

STRASSER, S., GOODMAN, R., SHEPPARD, J. and BUTCHER, S. (2016). A new discrete particle swarm optimization algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference* 2016 53–60. ACM, Denver, CO, USA. https://doi.org/10.1145/2908812.2908935

VEDALDI, A. and ZISSERMAN, A. (2012). Efficient additive kernels via explicit feature maps. *IEEE Trans. Pattern Anal. Mach. Intell.* **34** 480–492. https://doi.org/10.1109/TPAMI.2011.153

WANG, X., YAN, Y., TANG, P., BAI, X. and LIU, W. (2018). Revisiting multiple instance neural networks. *Pattern Recognit.* **74** 15–24. https://doi.org/10.1016/j.patcog.2017.08.026

WILLIAMS, C. and SEEGER, M. (2001). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems* **13** 682–688. MIT Press, Cambridge.

YANG, T., LI, Y.-F., MAHDAVI, M., JIN, R. and ZHOU, Z.-H. (2012). Nyström method vs random Fourier features: A theoretical and empirical comparison. *Adv. Neural Inf. Process. Syst.* **25** 476–484.

YUILLE, A. L. and RANGARAJAN, A. (2003). The concave-convex procedure. *Neural Comput.* **15** 915–936. Publisher: MIT Press. https://doi.org/10.1162/08997660360581958

ZELTZ, C., PRIMAC, I., ERUSAPPAN, P., ALAM, J., NOEL, A. and GULLBERG, D. (2020). Cancer-associated fibroblasts in desmoplastic tumors: Emerging role of integrins. *Semin. Cancer Biol.* **62** 166–181. https://doi.org/10.1016/j.semcancer.2019.08.004

ZHANG, H. H., AHN, J., LIN, X. and PARK, C. (2006). Gene selection using support vector machines with non-convex penalty. *Bioinformatics* **22** 88–95. Publisher: Oxford Univ. Press. MR2408601 https://doi.org/10.1016/j.csda.2007.02.006

ZHANG, X., WU, Y., WANG, L. and LI, R. (2016). Variable selection for support vector machines in moderately high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 53–76. MR3453646 https://doi.org/10.1111/rssb.12100

ZHAO, J., XU, Y., XU, C. and WANG, T. (2021). A two-stage safe screening method for non-convex support vector machine with ramp loss. *Knowl.-Based Syst.* **228** 107250. Publisher: Elsevier. MR3990392

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2207–2229
https://doi.org/10.1214/24-AOAS1877

PROBABILISTIC CONTRASTIVE DIMENSION REDUCTION FOR CASE-CONTROL STUDY DATA

BY DIDONG LI[1,a], ANDREW JONES[2,b] AND BARBARA ENGELHARDT[3,c]

[1]*Department of Biostatistics, University or North Carolina at Chapel Hill,* [a]*didongli@unc.edu*
[2]*Department of Computer Science, Princeton University,* [b]*aj13@princeton.edu*
[3]*Gladstone Institutes,* [c]*bengelhardt@stanford.edu*

Case-control experiments are essential to the scientific method, as they allow researchers to test biological hypotheses by looking for differences in outcome between cases and controls. It is then of interest to characterize variation that is enriched in a "foreground" (case) dataset relative to a "background" (control) dataset. For example, in a genomics context, the goal is to identify low-dimensional transcriptional structure unique to patients with certain disease (cases) vs. those without that disease (controls). In this work we propose probabilistic contrastive principal component analysis (PCPCA), a probabilistic dimension reduction method designed for case-control data. We describe inference in PCPCA through a contrastive likelihood and show that our model generalizes PCA, probabilistic PCA, and contrastive PCA. We discuss how to set the tuning parameter in theory and in practice, and we show several of PCPCA's advantages in the analysis of case-control data over related methods, including greater interpretability, uncertainty quantification and principled inference, robustness to noise and missing data, and the ability to generate "foreground-enriched" data from the model. We demonstrate PCPCA's performance on case-control data through a series of simulations, and we successfully identify variation specific to case data in genomic case-control experiments with data modalities, including gene expression, protein expression, and images.

REFERENCES

ABID, A., ZHANG, M. J., BAGARIA, V. K. and ZOU, J. (2017). Contrastive principal component analysis. arXiv preprint. Available at arXiv:1709.06716.

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.

ANDERSON, T. W. (1962). An introduction to multivariate statistical analysis. Technical report, Wiley New York. MR0091588

BHATTACHARYA, I. and MARTIN, R. (2022). Gibbs posterior inference on multivariate quantiles. *J. Statist. Plann. Inference* **218** 106–121. MR4337837 https://doi.org/10.1016/j.jspi.2021.10.003

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

BRENNER, N., BIALEK, W. and VAN STEVENINCK, R. D. R. (2000). Adaptive rescaling maximizes information transmission. *Neuron* **26** 695–702.

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

CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37. MR2811000 https://doi.org/10.1145/1970392.1970395

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M. A., GUO, J., LI, P. et al. (2017). Stan: A probabilistic programming language. *Grantee Submission* **76** 1–32.

Key words and phrases. High-dimensional data, contrastive model, probabilistic principal components analysis.

DARBYSHIRE, J. and HAMISH, J. (2016). *The Pricing and Hedging of Interest Rate Derivatives*: *A Practical Guide to Swaps*.

FRUCHTER, B. (1954). *Introduction to Factor Analysis*. Van Nostrand, Princeton.

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

GUAN, Y. and DY, J. (2009). Sparse probabilistic principal component analysis. In *Artificial Intelligence and Statistics* 185–192. PMLR.

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

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

HIGUERA, C., GARDINER, K. J. and CIOS, K. J. (2015). Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. *PLoS ONE* **10** e0129126.

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

HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24** 417.

IZENMAN, A. J. (2013). Linear discriminant analysis. In *Modern Multivariate Statistical Techniques* 237–280 Springer, New York.

JIANG, W. and TANNER, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.* **36** 2207–2231. MR2458185 https://doi.org/10.1214/07-AOS547

JIRSA, V. K., FRIEDRICH, R., HAKEN, H. and KELSO, J. S. (1994). A theoretical model of phase transitions in the human brain. *Biol. Cybernet.* **71** 27–35.

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

LAWRENCE, N. (2003). Gaussian process latent variable models for visualisation of high dimensional data. *Adv. Neural Inf. Process. Syst.* **16** 329–336.

LI, D., JONES, A. and ENGELHARDT, B. (2024). Supplement to "Probabilistic contrastive dimension reduction for case-control study data." https://doi.org/10.1214/24-AOAS1877SUPPA, https://doi.org/10.1214/24-AOAS1877SUPPB

LYDDON, S. P., HOLMES, C. C. and WALKER, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika* **106** 465–478. MR3949315 https://doi.org/10.1093/biomet/asz006

MATTEI, P.-A., BOUVEYRON, C. and LATOUCHE, P. (2016). Globally sparse probabilistic PCA. In *Artificial Intelligence and Statistics* 976–984. PMLR.

NOVEMBRE, J. and STEPHENS, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40** 646–649. https://doi.org/10.1038/ng.139

PASINI, G. (2017). Principal component analysis for stock portfolio management. *Int. J. Pure Appl. Math.* **115** 153–167.

QIAO, H. (2019). Discriminative principal component analysis: A reverse thinking. arXiv preprint. Available at arXiv:1903.04963.

RIGON, T., HERRING, A. H. and DUNSON, D. B. (2023). A generalized Bayes framework for probabilistic clustering. *Biometrika* **110** 559–578. MR4627771 https://doi.org/10.1093/biomet/asad004

ROUSSEEUW, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20** 53–65.

ROWE, D. B. (2003). *Multivariate Bayesian Statistics*: *Models for Source Separation and Signal Unmixing*. CRC Press, Boca Raton, FL. MR2000406

ROWEIS, S. T. (1998). EM algorithms for PCA and SPCA. *Adv. Neural Inf. Process. Syst.* 626–632.

RUSSAKOVSKY, O., DENG, J., SU, H. et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115** 211–252. MR3422482 https://doi.org/10.1007/s11263-015-0816-y

SCHÖLKOPF, B., SMOLA, A. and MÜLLER, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* **10** 1299–1319.

SEVERSON, K. A., GHOSH, S. and NG, K. (2019). Unsupervised learning with contrastive latent variable models. *Proc. AAAI Conf. Artif. Intell.* **33** 4862–4869.

SYRING, N. A. (2017). Gibbs posterior distributions: New theory and applications. PhD thesis, Univ. Illinois at Chicago. MR3828667

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

TWINE, N. A., JANITZ, K., WILKINS, M. R. and JANITZ, M. (2011). Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. *PLoS ONE* **6** e16266. https://doi.org/10.1371/journal.pone.0016266

VIDAL, R., MA, Y. and SASTRY, S. S. (2005). Generalized principal component analysis (GPCA). *IEEE Trans. Pattern Anal. Mach. Intell.* **27** 1945–1959.

WEINBERGER, E., BEEBE-WANG, N. and LEE, S.-I. (2022). Moment matching deep contrastive latent variable models. In *International Conference on Artificial Intelligence and Statistics* 2354–2371. PMLR.

WELLECK, S., KULIKOV, I., ROLLER, S., DINAN, E., CHO, K. and WESTON, J. (2019). Neural text generation with unlikelihood training. arXiv preprint. Available at arXiv:1908.04319.

WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534. https://doi.org/10.1093/biostatistics/kxp008

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.

ZHENG, G. X., TERRY, J. M., BELGRADER, P., RYVKIN, P., BENT, Z. W., WILSON, R., ZIRALDO, S. B., WHEELER, T. D., MCDERMOTT, G. P. et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8** 1–12.

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

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
2024, Vol. 18, No. 3, 2230–2253
https://doi.org/10.1214/24-AOAS1878

MODELING CURVES AND DERIVATIVES AS PREDICTORS FOR TRAFFIC BREAKDOWN PROBABILITIES

BY JENG-MIN CHIOU[1,a] AND PAI-LING LI[2,b]

[1]*Institute of Statistics and Data Science, National Taiwan University,* [a]*jmchiou@ntu.edu.tw*
[2]*Department of Statistics, Tamkang University,* [b]*plli@gms.tku.edu.tw*

Motivated by an interest in predicting the status of road traffic congestion within a short period, this paper presents a generalized functional linear regression model for predicting traffic breakdown probabilities. In this model, traffic congestion status is the response variable, and we utilize the observed traffic speed trajectories and their first two derivatives as functional predictors, representing different features of a random function. While the derivatives of a trajectory may contain useful information, they cannot be observed directly and so must be estimated. To address this challenge, we apply the Karhunen–Loève representation to individual functional predictors, including the trajectory and its derivatives. The regression model is reparameterized to represent both the integrated regression effect and the predictor-specific effects. The importance of these effects is indicated by the corresponding weight parameters. We also provide the consistency properties of the estimators relating to the derivative functional principal components and the regression parameter functions. In our simulation study, we find that the modeling approach is useful in its application to freeway traffic data; in particular, the use of speed trajectory derivatives as predictors for traffic status successfully enhances prediction accuracy.

REFERENCES

AHMEDOU, A., MARION, J.-M. and PUMO, B. (2016). Generalized linear model with functional predictors and their derivatives. *J. Multivariate Anal.* **146** 313–324. MR3477668 https://doi.org/10.1016/j.jmva.2015.10.009

ASH, R. B. and GARDNER, M. F. (1975). *Topics in Stochastic Processes*. *Probability and Mathematical Statistics*, *Vol.* 27. Academic Press, New York-London. MR0448463

CARDOT, H. and SARDA, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *J. Multivariate Anal.* **92** 24–41. MR2102242 https://doi.org/10.1016/j.jmva.2003.08.008

CARLSON, R. C., PAPAMICHAIL, I., PAPAGEORGIOU, M. and MESSMER, A. (2010). Optimal mainstream traffic flow control of large-scale motorway networks. *Transp. Res., Part C* **18** 193–212.

CHEN, K. and MÜLLER, H.-G. (2012). Conditional quantile analysis when covariates are functions, with application to growth data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 67–89. MR2885840 https://doi.org/10.1111/j.1467-9868.2011.01008.x

CHIOU, J.-M. (2012). Dynamical functional prediction and classification, with application to traffic flow prediction. *Ann. Appl. Stat.* **6** 1588–1614. MR3058676 https://doi.org/10.1214/12-AOAS595

CHIOU, J.-M. and LI, P.-L. (2024). Supplement to "Modeling curves and derivatives as predictors for traffic breakdown probabilities." https://doi.org/10.1214/24-AOAS1878SUPPA, https://doi.org/10.1214/24-AOAS1878SUPPB

CHIOU, J.-M., LIOU, H.-T. and CHEN, W.-H. (2021). Modeling time-varying variability and reliability of freeway travel time using functional principal component analysis. *IEEE Trans. Intell. Transp. Syst.* **22** 257–266.

CHIOU, J.-M. and MÜLLER, H.-G. (2009). Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *J. Amer. Statist. Assoc.* **104** 572–585. MR2751439 https://doi.org/10.1198/jasa.2009.0023

CHIOU, J.-M., MÜLLER, H.-G. and WANG, J.-L. (2003). Functional quasi-likelihood regression models with smooth random effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 405–423. MR1983755 https://doi.org/10.1111/1467-9868.00393

Key words and phrases. Derivatives, functional data, Karhunen–Loève representation, regression analysis, traffic congestion.

CHIOU, J.-M., ZHANG, Y.-C., CHEN, W.-H. and CHANG, C.-W. (2014). A functional data approach to missing value imputation and outlier detection for traffic flow data. *Transportmetrica B*: *Transp*. *Dyn*. **2** 106–129.

DAI, X., MÜLLER, H.-G. and TAO, W. (2018). Derivative principal component analysis for representing the time dynamics of longitudinal and functional data. *Statist*. *Sinica* **28** 1583–1609. MR3821019

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. CRC Press, London. MR1383587

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

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

HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis*, *with an Introduction to Linear Operators*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3379106 https://doi.org/10.1002/9781118762547

JAMES, G. M. (2002). Generalized linear models with functional predictors. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **64** 411–432. MR1924298 https://doi.org/10.1111/1467-9868.00342

LI, P.-L. and CHIOU, J.-M. (2021). Functional clustering and missing value imputation of traffic flow trajectories. *Transportmetrica B*: *Transp*. *Dyn*. **9** 1–21.

LIU, B. and MÜLLER, H.-G. (2009). Estimating derivatives for samples of sparsely observed functions, with application to online auction dynamics. *J*. *Amer*. *Statist*. *Assoc*. **104** 704–717. MR2541589 https://doi.org/10.1198/jasa.2009.0115

MAS, A. and PUMO, B. (2009). Functional linear regression with derivatives. *J*. *Nonparametr*. *Stat*. **21** 19–40. MR2483857 https://doi.org/10.1080/10485250802401046

MAY, A. D. (1990). *Traffic Flow Fundamentals*. Prentice Hall, Englewood Cliffs, NJ.

MORRIS, J. S. (2015). Functional regression. *Annu*. *Rev*. *Stat*. *Appl*. **2** 321–359.

MÜLLER, H.-G. and STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann*. *Statist*. **33** 774–805. MR2163159 https://doi.org/10.1214/009053604000001156

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

PERSAUD, B., YAGAR, S. and BROWNLEE, R. (1998). Exploration of the breakdown phenomenon in freeway traffic. *Transp*. *Res*. *Rec*. **1634** 64–69.

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

RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **53** 233–243. MR1094283

WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional data analysis. *Annu*. *Rev*. *Stat*. *Appl*. **3** 257–295.

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61** 439–447. MR0375592 https://doi.org/10.1093/biomet/61.3.439

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

ZHANG, J. T. (2013). *Analysis of Variance for Functional Data*. *Chapman & Hall/CRC Monographs on Statistics & Applied Probability*. Taylor & Francis, London.

ZHANG, X. and WANG, J.-L. (2016). From sparse to dense functional data and beyond. *Ann*. *Statist*. **44** 2281–2321. MR3546451 https://doi.org/10.1214/16-AOS1446

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2254–2276
https://doi.org/10.1214/24-AOAS1879

SURROGATE METHOD FOR PARTIAL ASSOCIATION BETWEEN MIXED DATA WITH APPLICATION TO WELL-BEING SURVEY ANALYSIS

BY SHAOBO LI[1,a], ZHAOHU FAN[2,b], IVY LIU[3,c], PHILIP S. MORRISON[4,d] AND DUNGANG LIU[5,e]

[1]*School of Business, University of Kansas,* [a]*shaobo.li@ku.edu*
[2]*Scheller College of Business, Georgia Institute of Technology,* [b]*jonathan.fan@scheller.gatech.edu*
[3]*School of Mathematics and Statistics, Victoria University of Wellington,* [c]*ivy.liu@vuw.ac.nz*
[4]*School of Geography, Environment and Earth Sciences, Victoria University of Wellington,* [d]*philip.morrison@vuw.ac.nz*
[5]*Lindner College of Business, University of Cincinnati,* [e]*dungang.liu@uc.edu*

This paper is motivated by the analysis of a survey study focusing on college student well-being before and after the COVID-19 pandemic outbreak. A statistical challenge in well-being studies lies in the multidimensionality of outcome variables, recorded in various scales such as continuous, binary, or ordinal. The presence of mixed data complicates the examination of their relationships when adjusting for important covariates. To address this challenge, we propose a unifying framework for studying partial association between mixed data. We achieve this by defining a unified residual using the surrogate method. The idea is to map the residual randomness to a consistent continuous scale, regardless of the original scales of outcome variables. This framework applies to parametric or semiparametric models for covariate adjustments. We validate the use of such residuals for assessing partial association, introducing a measure that generalizes classical Kendall's tau to capture both partial and marginal associations. Moreover, our development advances the theory of the surrogate method by demonstrating its applicability without requiring outcome variables to have a latent variable structure. In the analysis of the college student well-being survey, our proposed method unveils the contingency of relationships between multidimensional well-being measures and micro personal risk factors (e.g., physical health, loneliness, and accommodation) as well as the macro disruption caused by COVID-19.

REFERENCES

AGNIEL, D. and CAI, T. (2017). Analysis of multiple diverse phenotypes via semiparametric canonical correlation analysis. *Biometrics* **73** 1254–1265. MR3744539 https://doi.org/10.1111/biom.12690

AGRESTI, A. (2010). *Analysis of Ordinal Categorical Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2742515 https://doi.org/10.1002/9780470594001

ANDERSON, J. A. (1984). Regression and ordered categorical variables. *J. Roy. Statist. Soc. Ser. B* **46** 1–30. MR0745211

BAI, H., ZHONG, Y., GAO, X. and XU, W. (2020). Multivariate mixed response model with pairwise composite-likelihood method. *Stats* **3** 203–220.

BURNS, D., DAGNALL, N. and HOLT, M. (2020). Assessing the impact of the Covid-19 pandemic on student wellbeing at universities in the United Kingdom: A conceptual analysis. *Frontiers in Education* **5** 582882.

CAO, J., WEI, J., ZHU, H., DUAN, Y., GENG, W., HONG, X., JIANG, J., ZHAO, X. and ZHU, B. (2020). A study of basic needs and psychological wellbeing of medical workers in the fever clinic of a tertiary general hospital in Beijing during the Covid-19 outbreak. *Psychother. Psychosom.* **89** 252–254.

CATALANO, P. J. and RYAN, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Amer. Statist. Assoc.* **87** 651–658.

CHAMBERS, J. M., CLEVELAND, W. S., KLEINER, B. and TUKEY, P. A. (2018). *Graphical Methods for Data Analysis*. CRC Press/CRC, Boca Raton.

Key words and phrases. Covariate adjustment, COVID-19 pandemic, Kendall's tau, mental health, moderation effect, partial correlation, surrogate residual.

CHENG, C., WANG, R. and ZHANG, H. (2021). Surrogate residuals for discrete choice models. *J. Comput. Graph. Statist.* **30** 67–77. MR4235965 https://doi.org/10.1080/10618600.2020.1775618

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

COX, D. R. and WERMUTH, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika* **79** 441–461. MR1187603 https://doi.org/10.1093/biomet/79.3.441

DE LEON, A. R. and CARRIÈRE, K. C. (2007). General mixed-data model: Extension of general location and grouped continuous models. *Canad. J. Statist.* **35** 533–548. MR2381398 https://doi.org/10.1002/cjs.5550350405

DE LEON, A. R. and CHOUGH, K. C. (2013). *Analysis of Mixed Data: Methods & Applications*. CRC Press/CRC, Boca Raton.

DE LEON, A. R. and WU, B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Stat. Med.* **30** 175–185. MR2758273 https://doi.org/10.1002/sim.4087

DIENER, E., EMMONS, R. A., LARSEN, R. J. and GRIFFIN, S. (1985). The satisfaction with life scale: A measure of life satisfaction. *Journal of Personality Assessment* **49** 71–75.

DUNN, P. K. and SMYTH, G. K. (1996). Randomized quantile residuals. *J. Comput. Graph. Statist.* **5** 236–244.

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

EVERY-PALMER, S., JENKINS, M., GENDALL, P., HOEK, J., BEAGLEHOLE, B., BELL, C., WILLIMAN, J., RAPSEY, C. and STANLEY, J. (2020). Psychological distress, anxiety, family violence, suicidality, and wellbeing in New Zealand during the Covid-19 lockdown: A cross-sectional study. *PLoS ONE* **15** e0241658. https://doi.org/10.1371/journal.pone.0241658

FAES, C., AERTS, M., MOLENBERGHS, G., GEYS, H., TEUNS, G. and BIJNENS, L. (2008). A high-dimensional joint model for longitudinal outcomes of different nature. *Stat. Med.* **27** 4408–4427. MR2528521 https://doi.org/10.1002/sim.3314

FAN, J., LIU, H., NING, Y. and ZOU, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 405–421. MR3611752 https://doi.org/10.1111/rssb.12168

FERNANDEZ, D., LIU, I. and COSTILLA, R. (2019). A method for ordinal outcomes: The ordered stereotype model. *Int. J. Methods Psychiatr. Res.* **28** e1801.

FISHER, R. A. (1924). The distribution of the partial correlation coefficient. *Metron* **3** 329–332.

FITZMAURICE, G. M. and LAIRD, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *J. Amer. Statist. Assoc.* **90** 845–852. MR1354003

GREENLAND, S. (2008). Invited commentary: Variable selection versus shrinkage in the control of multiple confounders. *Amer. J. Epidemiol.* **167** 523–529.

GREENWELL, B. M., MCCARTHY, A. J., BOEHMKE, B. C. and LIU, D. (2018). Residuals and diagnostics for binary and ordinal regression models: An introduction to the sure package. *R J.* **10** 381–394.

GROARKE, J. M., BERRY, E., GRAHAM-WISENER, L., MCKENNA-PLUMLEY, P. E., MCGLINCHEY, E. and ARMOUR, C. (2020). Loneliness in the UK during the Covid-19 pandemic: Cross-sectional results from the Covid-19 Psychological Wellbeing Study. *PLoS ONE* **15** e0239698. https://doi.org/10.1371/journal.pone.0239698

GUEORGUIEVA, R. V. and AGRESTI, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous reponses. *J. Amer. Statist. Assoc.* **96** 1102–1112. MR1947258 https://doi.org/10.1198/016214501753208762

HE, J., LI, H., EDMONDSON, A. C., RADER, D. J. and LI, M. (2012). A Gaussian copula approach for the analysis of secondary phenotypes in case–control genetic association studies. *Biostatistics* **13** 497–508.

HONG, H. G. and HE, X. (2010). Prediction of functional status for the elderly based on a new ordinal regression model. *J. Amer. Statist. Assoc.* **105** 930–941. MR2752590 https://doi.org/10.1198/jasa.2010.ap08631

JIANG, Y., LI, N. and ZHANG, H. (2014). Identifying genetic variants for addiction via propensity score adjusted generalized Kendall's tau. *J. Amer. Statist. Assoc.* **109** 905–930. MR3265665 https://doi.org/10.1080/01621459.2014.901223

JOHNSON, R. A. and WICHERN, D. W. (2007). *Applied Multivariate Statistical Analysis*, 6th ed. Pearson Prentice Hall, Upper Saddle River, NJ. MR2372475

KAHNEMAN, D. (2011). *Thinking, Fast and Slow*. Macmillan, London.

KENDALL, M. G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–93.

KOPEC, J. A. and ESDAILE, J. M. (1990). Bias in case-control studies. A review. *J. Epidemiol. Community Health* **44** 179–186.

KOSMIDIS, I. (2021). Mean and median bias reduction: A concise review and application to adjacent-categories logit models. ArXiv preprint. Available at arXiv:2112.02621.

LEDERER, A. M., HOBAN, M. T., LIPSON, S. K., ZHOU, S. and EISENBERG, D. (2021). More than inconvenienced: The unique needs of US college students during the Covid-19 pandemic. *Health Educ. Behav.* **48** 14–19.

LEE, P. H. (2014). Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification? *J. Epidemiol.* **24** 161–167. https://doi.org/10.2188/jea.je20130062

LI, C. and SHEPHERD, B. E. (2010). Test of association between two ordinal variables while adjusting for covariates. *J. Amer. Statist. Assoc.* **105** 612–620. MR2724846 https://doi.org/10.1198/jasa.2010.tm09386

LI, C. and SHEPHERD, B. E. (2012). A new residual for ordinal outcomes. *Biometrika* **99** 473–480. MR2931266 https://doi.org/10.1093/biomet/asr073

LI, S., FAN, Z., LIU, I., MORRISON, P. S and LIU, D. (2024). Supplement to "Surrogate method for partial association between mixed data with application to well-being survey analysis." https://doi.org/10.1214/24-AOAS1879SUPP

LI, S., ZHU, X., CHEN, Y. and LIU, D. (2021). PAsso: An R package for assessing partial association between ordinal variables. *R J.* **13** 239–252.

LIU, D., LI, S., YU, Y. and MOUSTAKI, I. (2021). Assessing partial association between ordinal variables: Quantification, visualization, and hypothesis testing. *J. Amer. Statist. Assoc.* **116** 955–968. MR4270036 https://doi.org/10.1080/01621459.2020.1796394

LIU, D. and ZHANG, H. (2018). Residuals and diagnostics for ordinal regression models: A surrogate approach. *J. Amer. Statist. Assoc.* **113** 845–854. MR3832231 https://doi.org/10.1080/01621459.2017.1292915

LIU, D., ZHU, X., GREENWELL, B. and LIN, Z. (2023). A new goodness-of-fit measure for probit models: Surrogate R^2. *Br. J. Math. Stat. Psychol.* **76** 192–210.

LIU, I. and AGRESTI, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments (with discussion). *TEST* **14** 1–73. MR2203424 https://doi.org/10.1007/BF02595397

LIU, Q., LI, C., WANGA, V. and SHEPHERD, B. E. (2018). Covariate-adjusted Spearman's rank correlation with probability-scale residuals. *Biometrics* **74** 595–605. MR3825346 https://doi.org/10.1111/biom.12812

LIU, Q., SHEPHERD, B. and LI, C. (2020). PResiduals: An R package for residual analysis using probability-scale residuals. *J. Stat. Softw.* **94** 1–27.

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

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

MORRISON, P. S., ROSSOUW, S. and GREYLING, T. (2021). The impact of exogenous shocks on national well-being. New Zealanders' reaction to Covid-19. *Applied Research in Quality of Life* 1–26.

NAJITA, J. S., LI, Y. and CATALANO, P. J. (2009). A novel application of a bivariate regression model for binary and continuous outcomes to studies of fetal toxicity. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 555–573. MR2750093 https://doi.org/10.1111/j.1467-9876.2009.00667.x

PRENTICE, R. L. and ZHAO, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47** 825–839. MR1141951 https://doi.org/10.2307/2532642

SALES, A., DROLET, R. and BONNEAU, I. (2001). Academic paths, ageing and the living conditions of students in the late 20th century. *Canadian Review of Sociology* **38** 167–188.

SAMMEL, M. D., RYAN, L. M. and LEGLER, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **59** 667–678.

SCHKADE, D. A. and KAHNEMAN, D. (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychol. Sci.* **9** 340–346.

SHEPHERD, B. E., LI, C. and LIU, Q. (2016b). Probability-scale residuals for continuous, discrete, and censored data. *Canad. J. Statist.* **44** 463–479. MR3574132 https://doi.org/10.1002/cjs.11302

SONG, P. X.-K., LI, M. and YUAN, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics* **65** 60–68. MR2665846 https://doi.org/10.1111/j.1541-0420.2008.01058.x

SPITZER, R. L., KROENKE, K., WILLIAMS, J. B. W. and LÖWE, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Arch. Intern. Med.* **166** 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

STÖBER, J., HONG, H. G., CZADO, C. and GHOSH, P. (2015). Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses. *Comput. Statist. Data Anal.* **88** 28–39. MR3332015 https://doi.org/10.1016/j.csda.2015.02.001

TEIXEIRA-PINTO, A. and NORMAND, S.-L. T. (2009). Correlated bivariate continuous and binary outcomes: Issues and applications. *Stat. Med.* **28** 1753–1773. MR2751596 https://doi.org/10.1002/sim.3588

TOPP, C. W., ØSTERGAARD, S. D., SØNDERGAARD, S. and BECH, P. (2015). The WHO-5 Well-Being Index: A systematic review of the literature. *Psychother. Psychosom.* **84** 167–176.

TOUSSAINT, A., HÜSING, P., GUMZ, A., WINGENFELD, K., HÄRTER, M., SCHRAMM, E. and LÖWE, B. (2020). Sensitivity to change and minimal clinically important difference of the 7-item Generalized Anxiety Disorder Questionnaire (GAD-7). *J. Affective Disorders* **265** 395–401. https://doi.org/10.1016/j.jad.2020.01.032

TUTZ, G. (2022). Ordinal regression: A review and a taxonomy of models. *Wiley Interdiscip. Rev.: Comput. Stat.* **14** Paper No. e1545. MR4396895 https://doi.org/10.1002/wics.1545

WITTKAMPF, K. A., NAEIJE, L., SCHENE, A. H., HUYSER, J. and VAN WEERT, H. C. (2007). Diagnostic accuracy of the mood module of the Patient Health Questionnaire: A systematic review. *Gen. Hosp. Psychiatry* **29** 388–395. https://doi.org/10.1016/j.genhosppsych.2007.06.004

YANG, L. (2022). Nonparametric copula estimation for mixed insurance claim data. *J. Bus. Econom. Statist.* **40** 537–546. MR4410880 https://doi.org/10.1080/07350015.2020.1835668

ZHANG, H., LIU, D., ZHAO, J. and BI, X. (2018). Modeling hybrid traits for comorbidity and genetic studies of alcohol and nicotine co-dependence. *Ann. Appl. Stat.* **12** 2359–2378. MR3875704 https://doi.org/10.1214/18-AOAS1156

ZHAO, L. P., PRENTICE, R. L. and SELF, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model. *J. Roy. Statist. Soc. Ser. B* **54** 805–811.

ZHU, W., JIANG, Y. and ZHANG, H. (2012). Nonparametric covariate-adjusted association tests based on the generalized Kendall's tau. *J. Amer. Statist. Assoc.* **107** 1–11. MR2949337 https://doi.org/10.1080/01621459.2011.643707

ZILKO, A. A. and KUROWICKA, D. (2016). Copula in a multivariate mixed discrete-continuous model. *Comput. Statist. Data Anal.* **103** 28–55. MR3522617 https://doi.org/10.1016/j.csda.2016.02.017

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2277–2294
https://doi.org/10.1214/24-AOAS1880

AN INTEGRATIVE NETWORK-BASED MEDIATION MODEL (NMM) TO ESTIMATE MULTIPLE GENETIC EFFECTS ON OUTCOMES MEDIATED BY FUNCTIONAL CONNECTIVITY

BY WEI DAI[a] AND HEPING ZHANG[b]

Department of Biostatistics, Yale University School of Public Health, [a]*wei.dai.wd278@yale.edu,* [b]*heping.zhang@yale.edu*

Functional connectivity of the brain, characterized by interconnected neural circuits across functional networks, is a cutting-edge feature in neuroimaging. It has the potential to mediate the effect of genetic variants on behavioral outcomes or diseases. Existing mediation analysis methods can evaluate the impact of genetics and brain structure/function on cognitive behavior or disorders, but they tend to be limited to single genetic variants or univariate mediators, without considering cumulative genetic effects and the complex matrix and group and network structures of functional connectivity. To address this gap, the paper presents an integrative network-based mediation model (NMM) that estimates the effect of multiple genetic variants on behavioral outcomes or diseases mediated by functional connectivity. The model incorporates group information of inter-regions at broad network level and imposes low-rank and sparse assumptions to reflect the complex structures of functional connectivity and selecting network mediators simultaneously. We adopt block coordinate descent algorithm to implement a fast and efficient solution to our model. Simulation results indicate the efficacy of the model in selecting active mediators and reducing bias in effect estimation. With application to the Human Connectome Project Youth Adult (HCP-YA) study of 493 young adults, two genetic variants (rs769448 and rs769449) on the *APOE4* gene are identified that lead to deficits in functional connectivity within visual networks and fluid intelligence.

REFERENCES

BETZEL, R. F., MEDAGLIA, J. D. and BASSETT, D. S. (2018). Diversity of meso-scale architecture in human and non-human connectomes. *Nat. Commun.* **9** 346. https://doi.org/10.1038/s41467-017-02681-z

BI, X., YANG, L., LI, T., WANG, B., ZHU, H. and ZHANG, H. (2017). Genome-wide mediation analysis of psychiatric and cognitive traits through imaging phenotypes. *Hum. Brain Mapp.* **38** 4088–4097. https://doi.org/10.1002/hbm.23650

CHANG, W.-C., FANG, Y.-Y., CHANG, H.-W., CHUANG, L.-Y., LIN, Y.-D., HOU, M.-F. and YANG, C.-H. (2014). Identifying association model for single-nucleotide polymorphisms of ORAI1 gene for breast cancer. *Cancer Cell Int.* **14** 1–6.

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

CHUNG, M. K. (2018). Statistical challenges of big brain network data. *Statist. Probab. Lett.* **136** 78–82. MR3806842 https://doi.org/10.1016/j.spl.2018.02.020

DAI, W., NOBLE, S. and SCHEINOST, D. (2022). The semi-constrained Network-Based Statistic (scNBS): Integrating local and global information for brain network inference. In *Medical Image Computing and Computer Assisted Intervention–MICCAI* 2022: 25*th International Conference*, *Singapore*, *September* 18–22, 2022, *Proceedings*, *Part I* 396–405. Springer, New York.

DAI, W. and ZHANG, H. (2024). Supplement to "An integrative network-based mediation model (NMM) to estimate multiple genetic effects on outcomes mediated by functional connectivity." https://doi.org/10.1214/24-AOAS1880SUPP

EAVANI, H., SATTERTHWAITE, T. D., FILIPOVYCH, R., GUR, R. E., GUR, R. C. and DAVATZIKOS, C. (2015). Identifying sparse connectivity patterns in the brain using resting-state fMRI. *NeuroImage* **105** 286–299. https://doi.org/10.1016/j.neuroimage.2014.09.058

Key words and phrases. Mediation analysis, matrix mediators, imaging genetic, brain connectome.

FINN, E. S., SHEN, X., SCHEINOST, D., ROSENBERG, M. D., HUANG, J., CHUN, M. M., PAPADEMETRIS, X. and CONSTABLE, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* **18** 1664–1671.

KOROLOGOU-LINDEN, R., BHATTA, L., BRUMPTON, B. M., HOWE, L. D., MILLARD, L. A. C., KOLARIC, K., BEN-SHLOMO, Y., WILLIAMS, D. M., SMITH, G. D. et al. (2022). The causes and consequences of Alzheimer's disease: Phenome-wide evidence from Mendelian randomization. *Nat. Commun.* **13** 4726. https://doi.org/10.1038/s41467-022-32183-6

LEONARDI, N., RICHIARDI, J., GSCHWIND, M., SIMIONI, S., ANNONI, J. M., SCHLUEP, M., VUILLEUMIER, P. and VAN DE VILLE, D. (2013). Principal components of functional connectivity: A new approach to study dynamic brain connectivity during rest. *NeuroImage* **83** 937–950. https://doi.org/10.1016/j.neuroimage.2013.07.019

LI, K., GUO, L., NIE, J., LI, G. and LIU, T. (2009). Review of methods for functional brain connectivity detection using fMRI. *Comput. Med. Imaging Graph.* **33** 131–139. https://doi.org/10.1016/j.compmedimag.2008.10.011

LOEB, K. R. and LOEB, L. A. (2000). Significance of multiple mutations in cancer. *Carcinogenesis* **21** 379–385. https://doi.org/10.1093/carcin/21.3.379

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

MEUNIER, D., LAMBIOTTE, R. and BULLMORE, E. T. (2010). Modular and hierarchically modular organization of brain networks. *Front. Neurosci.* **4** 200. https://doi.org/10.3389/fnins.2010.00200.

NEWMAN, M. E. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577–8582. https://doi.org/10.1073/pnas.0601602103

NG, B., VAROQUAUX, G., POLINE, J. B. and THIRION, B. (2012). A novel sparse graphical approach for multimodal brain connectivity inference. *Med. Image Comput. Comput. Assist. Interv.* **15** 707–714.

NOBLE, S., MEJIA, A. F., ZALESKY, A. and SCHEINOST, D. (2022). Improving power in functional magnetic resonance imaging by moving beyond cluster-level inference. *Proc. Natl. Acad. Sci. USA* **119** e2203020119. https://doi.org/10.1073/pnas.2203020119

PASSAMONTI, L., RICCELLI, R., INDOVINA, I., DUGGENTO, A., TERRACCIANO, A. and TOSCHI, N. (2019). Time-resolved connectome of the five-factor model of personality. *Sci. Rep.* **9** 15066. https://doi.org/10.1038/s41598-019-51469-2

RODRIGUEZ, R. X., NOBLE, S., TEJAVIBULYA, L. and SCHEINOST, D. (2022). Leveraging edge-centric networks complements existing network-level inference for functional connectomes. *NeuroImage* **264** 119742. https://doi.org/10.1016/j.neuroimage.2022.119742

RONAN, L., VOETS, N. L., HOUGH, M., MACKAY, C., ROBERTS, N., SUCKLING, J., BULLMORE, E., JAMES, A. and FLETCHER, P. C. (2012). Consistency and interpretation of changes in millimeter-scale cortical intrinsic curvature across three independent datasets in schizophrenia. *NeuroImage* **63** 611–621. https://doi.org/10.1016/j.neuroimage.2012.06.034

ROSENBERG, M. D., FINN, E. S., SCHEINOST, D., PAPADEMETRIS, X., SHEN, X., CONSTABLE, R. T. and CHUN, M. M. (2016). A neuromarker of sustained attention from whole-brain functional connectivity. *Nat. Neurosci.* **19** 165–171.

RUBINOV, M. and SPORNS, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage* **52** 1059–1069. https://doi.org/10.1016/j.neuroimage.2009.10.003

SAFIEH, M., KORCZYN, A. D. and MICHAELSON, D. M. (2019). ApoE4: An emerging therapeutic target for Alzheimer's disease. *BMC Med.* **17** 1–17.

SALAS, N., ESCOBAR, J. and HUEPE, D. (2021). Two sides of the same coin: Fluid intelligence and crystallized intelligence as cognitive reserve predictors of social cognition and executive functions among vulnerable elderly people. *Front. Neurol.* **12** 599378. https://doi.org/10.3389/fneur.2021.599378

SHEN, X., FINN, E. S., SCHEINOST, D., ROSENBERG, M. D., CHUN, M. M., PAPADEMETRIS, X. and CONSTABLE, R. T. (2017). Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat. Protoc.* **12** 506–518. https://doi.org/10.1038/nprot.2016.178

SHEN, X., TOKOGLU, F., PAPADEMETRIS, X. and CONSTABLE, R. T. (2013). Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *NeuroImage* **82** 403–415. https://doi.org/10.1016/j.neuroimage.2013.05.081

SPORNS, O. and BETZEL, R. F. (2016). Modular brain networks. *Annu. Rev. Psychol.* **67** 613–640. https://doi.org/10.1146/annurev-psych-122414-033634

TINGLEY, D., YAMAMOTO, T., HIROSE, K., KEELE, L. and IMAI, K. (2014). Mediation: R package for causal mediation analysis.

TOST, H., BILEK, E. and MEYER-LINDENBERG, A. (2012). Brain connectivity in psychiatric imaging genetics. *NeuroImage* **62** 2250–2260. https://doi.org/10.1016/j.neuroimage.2011.11.007

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K. and Consortium, W.-M. H. (2013). The Wu-minn human connectome project: An overview. *NeuroImage* **80** 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

van Wijk, B. C. M., Stam, C. J. and Daffertshofer, A. (2010). Comparing brain networks of different size and connectivity density using graph theory. *PLoS ONE* **5** e13701. https://doi.org/10.1371/journal.pone.0013701

Woo, M. and Kim, Y. (2017). Cortical functional connections and fluid intelligence in adolescent APOE4 carriers. *Dement. Geriatr. Cogn. Disord.* **44** 153–159. https://doi.org/10.1159/000479276.

Xia, C. H., Ma, Z., Ciric, R., Gu, S., Betzel, R. F., Kaczkurkin, A. N., Calkins, M. E., Cook, P. A., de la Garza, A. G. et al. (2018). Linked dimensions of psychopathology and connectivity in functional brain networks. *Nat. Commun.* **9** 3003. https://doi.org/10.1038/s41467-018-05317-y

Xia, C. H., Ma, Z., Cui, Z., Bzdok, D., Thirion, B., Bassett, D. S., Satterthwaite, T. D., Shinohara, R. T. and Witten, D. M. (2020). Multi-scale network regression for brain-phenotype associations. *Hum. Brain Mapp.* **41** 2553–2566. https://doi.org/10.1002/hbm.24982

Yu, D., Wang, L., Kong, D. and Zhu, H. (2022). Mapping the genetic-imaging-clinical pathway with applications to Alzheimer's disease. *J. Amer. Statist. Assoc.* **117** 1656–1668. MR4528461 https://doi.org/10.1080/01621459.2022.2087658

Zhang, Q. (2022). High-dimensional mediation analysis with applications to causal gene identification. *Stat. Biosci.* **14** 432–451.

Zhao, B., Li, T., Smith, S. M., Xiong, D., Wang, X., Yang, Y., Luo, T., Zhu, Z., Shan, Y. et al. (2022a). Common variants contribute to intrinsic human brain functional networks. *Nat. Genet.* **54** 508–517. https://doi.org/10.1038/s41588-022-01039-6

Zhao, Y., Chen, T., Cai, J., Lichenstein, S., Potenza, M. N. and Yip, S. W. (2022b). Bayesian network mediation analysis with application to the brain functional connectome. *Stat. Med.* **41** 3991–4005. MR4474168 https://doi.org/10.1002/sim.9488

Zhao, Y. and Luo, X. (2016). Pathway lasso: estimate and select sparse mediation pathways with high dimensional mediators. arXiv preprint. Available at arXiv:1603.07749.

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2295–2306
https://doi.org/10.1214/24-AOAS1881
© Institute of Mathematical Statistics, 2024

SEMIPARAMETRIC LINEAR REGRESSION WITH AN INTERVAL-CENSORED COVARIATE IN THE ATHEROSCLEROSIS RISK IN COMMUNITIES STUDY

BY RICHARD SIZELOVE[1,a], DONGLIN ZENG[2,c] AND DAN-YU LIN[1,b]

[1]*Department of Biostatistics, University of North Carolina at Chapel Hill,* [a]*sizelove@live.unc.edu,* [b]*lin@bios.unc.edu*
[2]*Department of Biostatistics, University of Michigan,* [c]*dzeng@umich.edu*

In longitudinal studies, investigators are often interested in understanding how the time since the occurrence of an intermediate event affects a future outcome. The intermediate event is often asymptomatic such that its occurrence is only known to lie in a time interval induced by periodic examinations. We propose a linear regression model that relates the time since the occurrence of the intermediate event to a continuous response at a future time point through a rectified linear unit activation function while formulating the distribution of the time to the occurrence of the intermediate event through the Cox proportional hazards model. We consider nonparametric maximum likelihood estimation with an arbitrary sequence of examination times for each subject. We present an EM algorithm that converges stably for arbitrary datasets. The resulting estimators of regression parameters are consistent, asymptotically normal, and asymptotically efficient. We assess the performance of the proposed methods through extensive simulation studies and provide an application to the Atherosclerosis Risk in Communities Study.

REFERENCES

AHN, S., LIM, J., PAIK, M. C., SACCO, R. L. and ELKIND, M. S. (2018). Cox model with interval-censored covariate in cohort studies. *Biom. J.* **60** 797–814.

BENJAMIN, L. A., BRYER, A., EMSLEY, H. C., KHOO, S., SOLOMON, T. and CONNOR, M. D. (2012). HIV infection and stroke: Current perspectives and future directions. *Lancet Neurol.* **11** 878–890.

ECHOUFFO-TCHEUGUI, J. B., ZHANG, S., FLORIDO, R., HAMO, C., PANKOW, J. S., MICHOS, E. D., GOLDBERG, R. B., NAMBI, V., GERSTENBLITH, G. et al. (2021). Duration of diabetes and incident heart failure: The Atherosclerosis Risk in Communities (ARIC) study. *JACC Heart Fail.* **9** 594–603.

GOGGINS, W. B., FINKELSTEIN, D. M. and ZASLAVSKY, A. M. (1999). Applying the Cox proportional hazards model when the change time of a binary time-varying covariate is interval censored. *Biometrics* **55** 445–451. https://doi.org/10.1111/j.0006-341x.1999.00445.x

GOMEZ, G., ESPINAL, A. and LAGAKOS, S. W. (2003). Inference for a linear regression model with an interval-censored covariate. *Stat. Med.* **22** 409–425.

HSU, C.-H., TAYLOR, J. M. G., MURRAY, S. and COMMENGES, D. (2007). Multiple imputation for interval censored data with auxiliary variables. *Stat. Med.* **26** 769–781. https://doi.org/10.1002/sim.2581

KOZAKOVA, M. and PALOMBO, C. (2016). Diabetes mellitus, arterial wall, and cardiovascular risk assessment. *Int. J. Environ. Res. Public Health* **13** 201–215.

LANGOHR, K. and GOMEZ, G. (2014). Estimation and residual analysis with R for a linear regression model with an interval-censored covariate. *Biom. J.* **56** 867–885.

LANGOHR, K., GOMEZ, G. and MUGA, R. (2004). A parametric survival model with an interval-censored covariate. *Stat. Med.* **23** 3159–3175.

MORRISON, D., LAEYENDECKER, O. and BROOKMEYER, R. (2021). Regression with interval-censored covariates: Application to cross-sectional incidence estimation. *Biometrics* **78** 1–14.

MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–465.

SIZELOVE, R., ZENG, D. and LIN, D.-Y. (2024). Supplement to “Semiparametric linear regression with an interval-censored covariate in the atherosclerosis risk in communities study.” https://doi.org/10.1214/24-AOAS1881SUPPA, https://doi.org/10.1214/24-AOAS1881SUPPB

Key words and phrases. Duration of exposure, intermediate event, nonparametric likelihood, proportional hazards, rectified linear unit, semiparametric efficiency.

YU, B., PULIT, S. L., HWANG, S.-J., BRODY, J. A., AMIN, N., AUER, P. L., BIS, J. C., BOERWINKLE, E., BURKE, G. L. et al. (2016). Rare exome sequence variants in CLCN6 reduce blood pressure levels and hypertension risk. *Circ. Cardiovasc. Genet.* **9** 64–70.

ZENG, D., MAO, L. and LIN, D. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika* **103** 253–271.

THE ARIC INVESTIGATORS (1989). The Atherosclerosis Risk in Communities (ARIC) study: Design and objectives. *Amer. J. Epidemiol.* **129** 687–702.

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2307–2325
https://doi.org/10.1214/24-AOAS1882

SEMIPARAMETRIC MODELING OF SARS-COV-2 TRANSMISSION USING TESTS, CASES, DEATHS, AND SEROPREVALENCE DATA

BY DAMON BAYER[1,a], ISAAC H. GOLDSTEIN[1,b], JONATHAN FINTZI[2,d], KEITH LUMBARD[3,e], EMILY RICOTTA[4,f], SARAH WARNER[5,g], JEFFREY R STRICH[5,h], DANIEL S. CHERTOW[5,i], LINDSAY M. BUSCH[6,j], DANIEL M. PARKER[7,k], BERNADETTE BODEN-ALBALA[7,l], RICHARD CHHUON[8,m], MATTHEW ZAHN[8,n], NICHOLE QUICK[9,o], ALISSA DRATCH[10,p] AND VOLODYMYR M. MININ[1,c]

[1]*Department of Statistics, University of California, Irvine,* [a]*bayerd@uci.edu,* [b]*igoldst1@uci.edu,* [c]*vminin@uci.edu*
[2]*Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases,* [d]*jon.fintzi@nih.gov*
[3]*Clinical Monitoring Research Program Directorate, Frederick National Laboratory for Cancer Research,* [e]*keith.lumbard@nih.gov*
[4]*Epidemiology Unit, National Institute of Allergy and Infectious Diseases,* [f]*emily.ricotta@nih.gov*
[5]*Critical Care Medicine Department, Clinical Center, National Institutes of Health,* [g]*sarah.warner@nih.gov,* [h]*jeffrey.strich@nih.gov,* [i]*chertowd@cc.nih.gov*
[6]*Division of Infectious Diseases, Emory University School of Medicine,* [j]*lindsay.margoles.busch@emory.edu*
[7]*Susan and Henry Samueli College of Health Sciences, University of California, Irvine,* [k]*dparker1@hs.uci.edu,* [l]*bbodenal@hs.uci.edu*
[8]*Orange County Health Care Agency,* [m]*RChhuon@ochca.com,* [n]*MZahn@ochca.com*
[9]*Los Angeles County Department of Public Health,* [o]*NQuick@ph.lacounty.gov*
[10]*Edwards Lifesciences,* [p]*alissa.dratch@gmail.com*

Mechanistic models fit to streaming surveillance data are critical for understanding the transmission dynamics of an outbreak as it unfolds in real-time. However, transmission model parameter estimation can be imprecise, sometimes even impossible, because surveillance data are noisy and not informative about all aspects of the mechanistic model. To partially overcome this obstacle, Bayesian models have been proposed to integrate multiple surveillance data streams. We devised a modeling framework for integrating SARS-CoV-2 diagnostics test and mortality time series data as well as seroprevalence data from cross-sectional studies and tested the importance of individual data streams for both inference and forecasting. Importantly, our model for incidence data accounts for changes in the total number of tests performed. We apply our Bayesian data integration method to COVID-19 surveillance data collected in Orange County, California, between March 2020 and February 2021 and find that 32–72% of the Orange County residents experienced SARS-CoV-2 infection by mid-January, 2021. Despite this high number of infections, our results suggest that the abrupt end of the winter surge in January 2021 was due to both behavioral changes and a high level of accumulated natural immunity.

REFERENCES

ABBOTT, S., HELLEWELL, J., THOMPSON, R. N., SHERRATT, K., GIBBS, H. P., BOSSE, N. I., MUNDAY, J. D., MEAKIN, S., DOUGHTY, E. L. et al. (2020). Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Res*. **5** 112.

ANDERSON, S. C., EDWARDS, A. M., YERLANOV, M., MULBERRY, N., STOCKDALE, J. E., IYANIWURA, S. A., FALCAO, R. C., OTTERSTATTER, M. C., IRVINE, M. A. et al. (2020). Quantifying the impact of COVID-19 control measures using a Bayesian model of physical distancing. *PLoS Comput*. *Biol*. **16** 1–15.

ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **72** 269–342. MR2758115 https://doi.org/10.1111/j.1467-9868.2009.00736.x

Key words and phrases. Epidemic models, SARS-CoV-2.

Bayer, D., Goldstein, I. H., Fintzi, J., Lumbard, K., Ricotta, E., Warner, S., Strich, J. R., Chertow, D. S., Busch, L. M., Parker, D. M., Boden-Albala, B., Chhuon, R., Zahn, M., Quick, N., Dratch, A. and Minin, V. M. (2024). Supplement to "Semiparametric modeling of SARS-CoV-2 transmission using tests, cases, deaths, and seroprevalence data." https://doi.org/10.1214/24-AOAS1882SUPPA, https://doi.org/10.1214/24-AOAS1882SUPPB

Bhargava, A., Fukushima, E. A., Levine, M., Zhao, W., Tanveer, F., Szpunar, S. M. and Saravolatz, L. (2020). Predictors for severe COVID-19 infection. *Clin. Infect. Dis.* **71** 1962–1968.

Bosse, N. I., Gruson, H., Cori, A., van Leeuwen, E., Funk, S. and Abbott, S. (2022). Evaluating forecasts with scoringutils in R. ArXiv preprint. Available at arXiv:2205.07090.

Bretó, C., He, D., Ionides, E. L. and King, A. A. (2009). Time series analysis via mechanistic models. *Ann. Appl. Stat.* **3** 319–348. MR2668710 https://doi.org/10.1214/08-AOAS201

Bretó, C. and Ionides, E. L. (2011). Compound Markov counting processes and their applications to modeling infinitesimally over-dispersed systems. *Stochastic Process. Appl.* **121** 2571–2591. MR2832414 https://doi.org/10.1016/j.spa.2011.07.005

Bruckner, T. A., Parker, D. M., Bartell, S. M., Vieira, V. M., Khan, S., Noymer, A., Drum, E., Albala, B., Zahn, M. et al. (2021). Estimated seroprevalence of SARS-CoV-2 antibodies among adults in Orange County, California. *Sci. Rep.* **11** 3081.

United States Census Bureau (2020). Quick Facts: Orange County, California. https://www.census.gov/quickfacts/orangecountycalifornia. Accessed: 2020-09-05.

Cascante-Vega, J., Cordovez, J. M. and Santos-Vega, M. (2022). Estimating and forecasting the burden and spread of Colombia's SARS-CoV-2 first wave. *Sci. Rep.* **12** 13568.

Cauchemez, S. and Ferguson, N. M. (2008). Likelihood-based estimation of continuous-time epidemic models from time-series data: Application to measles transmission in London. *J. R. Soc. Interface* **5** 885–897.

Cori, A., Ferguson, N. M., Fraser, C. and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *Amer. J. Epidemiol.* **178** 1505–1512.

Cummings, M. J., Baldwin, M. R., Abrams, D., Jacobson, S. D., Meyer, B. J., Balough, E. M., Aaron, J. G., Claassen, J., Rabbani, L. E. et al. (2020). Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: A prospective cohort study. *Lancet* **395** 1763–1770.

Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A. B., Russell, T. W., Tully, D. C. et al. (2021). Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**.

Davies, N. G., Kucharski, A. J., Eggo, R. M., Gimma, A., Edmunds, W. J., Jombart, T., O'Reilly, K., Endo, A., Hellewell, J. et al. (2020). Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: A modelling study. *Lancet Public Health* **5** e375–e385.

Dukic, V., Lopes, H. F. and Polson, N. G. (2012). Tracking epidemics with Google Flu Trends data and a state-space SEIR model. *J. Amer. Statist. Assoc.* **107** 1410–1426. MR3036404 https://doi.org/10.1080/01621459.2012.713876

Ferguson, N. M. et al. (2020). Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. MRC Centre for Global Infectious Disease Analysis Reports. Accessed: 2020-06-19.

Fintzi, J., Wakefield, J. and Minin, V. N. (2022). A linear noise approximation for stochastic epidemic models fit to partially observed incidence counts. *Biometrics* **78** 1530–1541. MR4534376 https://doi.org/10.1111/biom.13538

Ge, H., Xu, K. and Ghahramani, Z. (2018). Turing: A language for flexible probabilistic inference. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (A. Storkey and F. Perez-Cruz, eds.). *Proceedings of Machine Learning Research* **84** 1682–1690. PMLR.

Gibson, G. C., Reich, N. G. and Sheldon, D. (2020). Real-time mechanistic Bayesian forecasts of COVID-19 mortality. *MedRxiv*.

Gleeson, J. P., Brendan Murphy, T., O'Brien, J. D., Friel, N., Bargary, N. and O'Sullivan, D. J. (2022). Calibrating COVID-19 susceptible-exposed-infected-removed models with time-varying effective contact rates. *Philos. Trans. R. Soc. Lond. A* **380** 20210120.

Grint, D. J., Wing, K., Houlihan, C., Gibbs, H. P., Evans, S. J., Williamson, E., McDonald, H. I., Bhaskaran, K., Evans, D. et al. (2022). Severity of SARS-CoV-2 alpha variant (B.1.1.7) in England. *Clin. Infect. Dis.* **75** e1120–e1127.

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

Höhle, M. and an der Heiden, M. (2014). Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics* **70** 993–1002. MR3295760 https://doi.org/10.1111/biom.12194

IRONS, N. J. and RAFTERY, A. E. (2021). Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys. *Proc. Natl. Acad. Sci. USA* **118** e2103272118.

JEWELL, S., FUTOMA, J., HANNAH, L., MILLER, A. C., FOTI, N. J. and FOX, E. B. (2021). It's complicated: Characterizing the time-varying relationship between cell phone mobility and COVID-19 spread in the US. *NPJ Digital Medicine* **4** 152.

JORDAN, A., KRÜGER, F. and LERCH, S. (2019). Evaluating probabilistic forecasts with scoringRules. *J. Stat. Softw.* **90** 1–37.

KIM, L., GARG, S., O'HALLORAN, A., WHITAKER, M., PHAM, H., ANDERSON, E. J., ARMISTEAD, I., BENNETT, N. M., BILLING, L. et al. (2021). Risk factors for intensive care unit admission and in-hospital mortality among hospitalized adults identified through the US coronavirus disease 2019 (COVID-19)-associated hospitalization surveillance network (COVID-NET). *Clin. Infect. Dis.* **72** e206–e214.

KNOCK, E. S., WHITTLES, L. K., LEES, J. A., PEREZ-GUZMAN, P. N., VERITY, R., FITZJOHN, R. G., GAYTHORPE, K. A. M., IMAI, N., HINSLEY, W. et al. (2021). Key epidemiological drivers and impact of interventions in the 2020 SARS-CoV-2 epidemic in England. *Sci. Transl. Med.* **13** eabg4262.

LEKONE, P. E. and FINKENSTÄDT, B. F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics* **62** 1170–1177. MR2307442 https://doi.org/10.1111/j.1541-0420.2006.00609.x

LI, J. and BRAUER, F. (2008). Continuous-time age-structured models in population dynamics and epidemiology. In *Mathematical Epidemiology* (F. Brauer, P. van den Driessche and J. Wu, eds.). *Lecture Notes in Math.* **1945** 205–227. Springer, Berlin. MR2428379 https://doi.org/10.1007/978-3-540-78911-6_9

MATHESON, J. E. and WINKLER, R. L. (1976). Scoring rules for continuous probability distributions. *Manage. Sci.* **22** 1087–1096.

MOROZOVA, O., LI, Z. R. and CRAWFORD, F. W. (2021). One year of modeling and forecasting COVID-19 transmission to support policymakers in Connecticut. *Sci. Rep.* **13** 20271.

NGUYEN-VAN-YEN, B., DEL MORAL, P. and CAZELLES, B. (2021). Stochastic epidemic models inference and diagnosis with Poisson random measure data augmentation. *Math. Biosci.* **335** Paper No. 108583. MR4236034 https://doi.org/10.1016/j.mbs.2021.108583

O'DEA, E. B. and DRAKE, J. M. (2022). A semi-parametric, state-space compartmental model with time-dependent parameters for forecasting COVID-19 cases, hospitalizations and deaths. *J. R. Soc. Interface* **19** 20210702.

OMORI, R., MIZUMOTO, K. and CHOWELL, G. (2020). Changes in testing rates could mask the novel coronavirus disease (COVID-19) growth rate. *Int. J. Infect. Dis.* **94** 116–118.

PEI, S., YAMANA, T. K., KANDULA, S., GALANTI, M. and SHAMAN, J. (2021). Burden and characteristics of COVID-19 in the United States during 2020. *Nature* 1–18.

PETRILLI, C. M., JONES, S. A., YANG, J., RAJAGOPALAN, H., O'DONNELL, L., CHERNYAK, Y., TOBIN, K. A., CERFOLIO, R. J., FRANCOIS, F. et al. (2020). Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: Prospective cohort study. *BMJ* **369**.

POOLEY, C. M., BISHOP, S. C. and MARION, G. (2015). Using model-based proposals for fast parameter inference on discrete state space, continuous-time Markov processes. *J. R. Soc. Interface* **12** 20150225.

PREM, K. et al. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: A modelling study. *Lancet Public Health* **5** e261–e270.

ROQUES, L., KLEIN, E. K., PAPAIX, J., SAR, A. and SOUBEYRAND, S. (2020). Using early data to estimate the actual infection fatality ratio from COVID-19 in France. *Biology* **9** 97.

SCOTT, J. A., GANDY, A., MISHRA, S., UNWIN, J., FLAXMAN, S. and BHATT, S. (2020). epidemia: Modeling of Epidemics using Hierarchical Bayesian Models. R package version 1.0.0.

SHUBIN, M., LEBEDEV, A., LYYTIKÄINEN, O. and AURANEN, K. (2016). Revealing the true incidence of pandemic A(H1N1)pdm09 influenza in Finland during the first two seasons—an analysis based on a dynamic transmission model. *PLoS Comput. Biol.* **12** 1–19.

SONG, J.-W., ZHANG, C., FAN, X., MENG, F.-P., XU, Z., XIA, P., CAO, W.-J., YANG, T., DAI, X.-P. et al. (2020). Immunological and inflammatory profiles in mild and severe cases of COVID-19. *Nat. Commun.* **11** 1–10.

STOKES, A. C., LUNDBERG, D. J., ELO, I. T., HEMPSTEAD, K., BOR, J. and PRESTON, S. H. (2021). COVID-19 and excess mortality in the United States: A county-level analysis. *PLoS Med.* **18** 1–18.

STONER, O. and ECONOMOU, T. (2020). Multivariate hierarchical frameworks for modeling delayed reporting in count data. *Biometrics* **76** 789–798. MR4151848 https://doi.org/10.1111/biom.13188

TEH, Y. W., ELESEDY, B., HE, B., HUTCHINSON, M., ZAIDI, S., BHOOPCHAND, A., PAQUET, U., TOMASEV, N., READ, J. et al. (2022). Efficient Bayesian inference of instantaneous reproduction numbers at fine spatial scales, with an application to mapping and nowcasting the Covid-19 epidemic in British local authorities. *J. Roy. Statist. Soc. Ser. A* **185** S65–S85. MR4547502

VAN DEN DRIESSCHE, P. (2008). Spatial structure: Patch models. In *Mathematical Epidemiology* (F. Brauer, P. van den Driessche and J. Wu, eds.). *Lecture Notes in Math.* **1945** 179–189. Springer, Berlin. MR2428377 https://doi.org/10.1007/978-3-540-78911-6_7

WHO (2021). Word Health Organization Q&A: Coronavirus disease (COVID-19): How is it transmitted? https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-covid-19-how-is-it-transmitted. Accessed: 2024-06-29.

WIKLE, N. B., TRAN, T. N.-A., GENTILESCO, B., LEIGHOW, S. M., ALBERT, E., STRONG, E. R., BRINDA, K., INAM, H., YANG, F. et al. (2022). SARS-CoV-2 epidemic after social and economic reopening in three US states reveals shifts in age structure and clinical characteristics. *Sci. Adv.* **8** eabf9868.

WU, Z. and MCGOOGAN, J. M. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *JAMA* **323** 1239–1242.

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2326–2341
https://doi.org/10.1214/24-AOAS1883

BAYESIAN JOINT MODELING OF HIGH-DIMENSIONAL DISCRETE MULTIVARIATE LONGITUDINAL DATA USING GENERALIZED LINEAR MIXED MODELS

BY PALOMA HAUSER[1,a], XIANMING TAN[1,b], FANG CHEN[2,d], RONALD C. CHEN[3,e] AND JOSEPH G. IBRAHIM[1,c]

[1]*Department of Biostatistics, University of North Carolina at Chapel Hill,* [a]*phauser@live.unc.edu,* [b]*xianming@email.unc.edu,* [c]*ibrahim@bios.unc.edu*

[2]*SAS Institute Inc.,* [d]*fangk.chen@sas.com*

[3]*Department of Radiation Oncology, University of Kansas Cancer Center,* [e]*rchen2@kumc.edu*

In routine cancer care, various patient- and clinician-reported symptoms are collected throughout treatment. This informs a crucial part of clinical research, particularly in studying the factors associated with symptom underascertainment. To jointly analyze such discrete, multivariate, and potentially high-dimensional repeated measures, we propose a Bayesian longitudinal generalized linear mixed model (BLGLMM). This model integrates three key methodologies: a low-rank matrix decomposition to approximate the high-dimensional regression coefficient matrix, a sparse factor model to capture the dependence among multiple outcomes, and random effects to account for the dependence among repeated responses. Posterior computation is performed using an efficient Markov chain Monte Carlo algorithm. We conduct simulations and provide an illustrative example examining the factors associated with symptom underascertainment in prostate cancer patients to demonstrate the efficacy and utility of our proposed method.

REFERENCES

BAI, J. and LI, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40** 436–465. MR3014313 https://doi.org/10.1214/11-AOS966

BASCH, E. (2014). The rationale for collecting patient-reported symptoms during routine chemotherapy. *Amer. Soc. Clin. Oncol. Educ. Book* 161–165. https://doi.org/10.14694/EdBook_AM.2014.34.161

BECKETT, L. A., TANCREDI, D. J. and WILSON, R. S. (2004). Multivariate longitudinal models for complex change processes. *Stat. Med.* **23** 231–239.

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. MR2806429 https://doi.org/10.1093/biomet/asr013

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

CHAKRABORTY, H., HELMS, R. W., SEN, P. K. and COHEN, M. S. (2003). Estimating correlation by using a general linear mixed model: Evaluation of the relationship between the concentration of HIV-1 RNA in blood and semen. *Stat. Med.* **22** 1457–1464. https://doi.org/10.1002/sim.1505

CHEN, K., CHAN, K.-S. and STENSETH, N. C. (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 203–221. MR2899860 https://doi.org/10.1111/j.1467-9868.2011.01002.x

CLARK, J. A. and TALCOTT, J. A. (2001). Symptom indexes to assess outcomes of treatment for early prostate cancer. *Med. Care* **39** 1118–1130. https://doi.org/10.1097/00005650-200110000-00009

DENIS, F., BASCH, E., SEPTANS, A.-L., BENNOUNA, J., URBAN, T., DUECK, A. C. and LETELLIER, C. (2019). Two-year survival comparing web-based symptom monitoring vs routine surveillance following treatment for lung cancer. *JAMA* **321** 306–307. https://doi.org/10.1001/jama.2018.18085

DIGGLE, P. J., HEAGERTY, P., LIANG, K. Y., HEAGERTY, P. and ZEGER, S. (2002). *Analysis of Longitudinal Data*, 1st ed. Oxford University Press, Oxford, UK.

Key words and phrases. Generalized linear mixed models, high-dimensional, longitudinal data, Markov chain Monte Carlo, low-rank approximation, patient-reported outcomes.

FAES, C., AERTS, M., MOLENBERGHS, G., GEYS, H., TEUNS, G. and BIJNENS, L. (2008). A high-dimensional joint model for longitudinal outcomes of different nature. *Stat. Med.* **27** 4408–4427. MR2528521 https://doi.org/10.1002/sim.3314

FIEUWS, S. and VERBEKE, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics* **62** 424–431. MR2227490 https://doi.org/10.1111/j.1541-0420.2006.00507.x

GAMERMAN, D. (1997). Sampling from the posterior distribution in generalized linear models. *Stat. Comput.* **7** 57–68.

GEWEKE, J. and ZHOU, G. (1996). Measuring the price of the arbitrage pricing theory. *Rev. Financ. Stud.* **9** 557–587.

HAUSER, P., TAN, X., CHEN, F., CHEN, R. C and IBRAHIM, J. G (2024). Supplement to "Bayesian joint modeling of high-dimensional discrete multivariate longitudinal data using generalized linear mixed models." https://doi.org/10.1214/24-AOAS1883SUPPA, https://doi.org/10.1214/24-AOAS1883SUPPB

HAUSER, P., TAN, X., CHEN, F. and IBRAHIM, J. G. (2023). Bayesian generalized linear low rank regression models for the detection of vaccine-adverse event associations. *Stat. Med.* **42** 2009–2026. MR4591609 https://doi.org/10.1002/sim.9711

HAVE, T. R. T. and MORABIA, A. (1999). Mixed effects models with bivariate and univariate association parameters for longitudinal bivariate binary response data. *Biometrics* **55** 85–93. https://doi.org/10.1111/j.0006-341x.1999.00085.x

HUGHES, D. M., GARCÍA-FIÑANA, M. and WAND, M. P. (2023). Fast approximate inference for multivariate longitudinal data. *Biostatistics* **24** 177–192. MR4522709 https://doi.org/10.1093/biostatistics/kxab021

HUI, F. K. C., MÜLLER, S. and WELSH, A. H. (2018). Sparse pairwise likelihood estimation for multivariate longitudinal mixed models. *J. Amer. Statist. Assoc.* **113** 1759–1769. MR3902244 https://doi.org/10.1080/01621459.2017.1371026

ILK, O. and DANIELS, M. J. (2007). Marginalized transition random effect models for multivariate longitudinal binary data. *Canad. J. Statist.* **35** 105–123. MR2345377 https://doi.org/10.1002/cjs.5550350110

IVANOVA, A., MOLENBERGHS, G. and VERBEKE, G. (2017). Fast and highly efficient pseudo-likelihood methodology for large and complex ordinal data. *Stat. Methods Med. Res.* **26** 2758–2779. MR3738281 https://doi.org/10.1177/0962280215608213

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

LU, Z. H., KHONDKER, Z., IBRAHIM, J. G., WANG, Y. and ZHU, H. (2017). Bayesian longitudinal low-rank regression models for imaging genetic data from longitudinal studies. *NeuroImage* **149** 305–322.

MAUFF, K., ERLER, N. S., KARDYS, I. and RIZOPOULOS, D. (2021). Pairwise estimation of multivariate longitudinal outcomes in a Bayesian setting with extensions to the joint model. *Stat. Model.* **21** 115–136. MR4209647 https://doi.org/10.1177/1471082X20945069

MCCULLOCH, C. E., SEARLE, S. R. and NEUHAUS, J. M. (2008). *Generalized, Linear, and Mixed Models*, 1st ed. Wiley, New York.

MOLENBERGHS, G. and VERBEKE, G. (2005). *Models for Discrete Longitudinal Data*, 1st ed. *Springer Series in Statistics*. Springer, New York. MR2171048

NEAL, P. and ROBERTS, G. (2006). Optimal scaling for partially updating MCMC algorithms. *Ann. Appl. Probab.* **16** 475–515. MR2244423 https://doi.org/10.1214/105051605000000791

PAKHOMOV, S., JACOBSEN, S., CHUTE, C. and ROGER, V. (2008). Agreement between patient-reported symptoms and their documentation in the medical record. *Amer. J. Manag. Care* **14** 530–539.

RIBAUDO, H. J. and THOMPSON, S. G. (2002). The analysis of repeated multivariate binary quality of life data: A hierarchical model approach. *Stat. Methods Med. Res.* **11** 69–83.

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

SUD, S., GERRINGER, B. C., WACASER, B. S., TAN, X., TATKOM, S. S., ROYCE, T. J., WANG, A. Z. and CHEN, R. C. (2021). Underascertainment of clinically meaningful symptoms during prostate cancer radiation therapy-does this vary by patient characteristics? *Int. J. Radiat. Oncol. Biol. Phys.* **110** 1122–1128.

US CENSUS BUREAU (2016). SAIPE state and county estimates for 2016: North Carolina. Available at https://www.census.gov/data/datasets/2016/demo/saipe/2016-state-and-county.html.

US DEPARTMENT OF AGRICULTURE, ECONOMIC RESEARCH SERVICE (2013). 2013 Rural-urban continuum codes. Available at https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx.

VERBEKE, G., FIEUWS, S., MOLENBERGHS, G. and DAVIDIAN, M. (2014). The analysis of multivariate longitudinal data: A review. *Stat. Methods Med. Res.* **23** 42–59. MR3190686 https://doi.org/10.1177/0962280212445834

VERBEKE, G. and MOLENBERGHS, G. (2000). *Linear Mixed Models for Longitudinal Data*, 1st ed. *Springer Series in Statistics*. Springer, New York. MR1880596 https://doi.org/10.1007/978-1-4419-0300-6

WANG, J. and LUO, S. (2017). Multidimensional latent trait linear mixed model: An application in clinical studies with multivariate longitudinal outcomes. *Stat. Med.* **36** 3244–3256. MR3689065 https://doi.org/10.1002/sim.7347

XIA, H. A., MA, H. and CARLIN, B. P. (2011). Bayesian hierarchical modeling for detecting safety signals in clinical trials. *J. Biopharm. Statist.* **21** 1006–1029. MR2823363 https://doi.org/10.1080/10543406.2010.520181

XIAO, C., POLOMANO, R. and BRUNER, D. W. (2013). Comparison between patient-reported and clinician-observed symptoms in oncology. *Cancer Nurs.* **36** E1–E16. https://doi.org/10.1097/NCC.0b013e318269040f

ZHU, H., KHONDKER, Z., LU, Z. and IBRAHIM, J. G. (2014). Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *J. Amer. Statist. Assoc.* **109** 977–990. MR3265670 https://doi.org/10.1080/01621459.2014.923775

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2342–2358
https://doi.org/10.1214/24-AOAS1884

ASSESSING AQUATIC TOXICITY ASSESSMENT VIA A CLUSTERED VARIANCE MODEL

BY XIN WANG[1,a] AND JING ZHANG[2,b]

[1]*Department of Mathematics and Statistics, San Diego State University,* [a]*xwang14@sdsu.edu*
[2]*Department of Statistics, Miami University,* [b]*zhangj8@miamioh.edu*

Motivated by the need to assess consistency in the outcomes of aquatic toxicity tests conducted by different labs at different time points, we propose a clustering of variance method in linear mixed models. The proposed method, referred as CVM, is able to identify the cluster structure of the variances and estimate model parameters simultaneously. In our proposed method, a penalized approach based on pairwise penalties is proposed to identify the cluster structure. We construct an optimization problem and develop an algorithm based on the alternating direction method of multipliers. Simulation studies show that the proposed approach can identify the cluster structure well and outperforms traditional methods based on k-means. In the end, the proposed approach is applied to the aquatic toxicity assessment data, which gives a more reasonable cluster structure than the traditional methods.

REFERENCES

AMATO, J. R., LUKASEWYCZ, M. T., ROBERT, E. D., MOUNT, D. I., DURHAN, E. J. and GERALD, T. A. (1993). An example of the identification of diazinon as a primary toxicant in an effluent. *Environ. Toxicol. Chem.* **11** 209–216.

ARCHAMBEAU, C., LEE, J. and VERLEYSEN, M. (2003). On convergence problems of the EM algorithm for finite Gaussian mixtures. In *European Symposium on Artificial Neural Networks* (*ESANN'*2003) 99–104, Bruges.

BAILER, A. J. and ORIS, J. T. (1993). Modeling reproductive toxicity in Ceriodaphnia tests. *Environ. Toxicol. Chem.* **12** 787–791.

BAILER, A. J. and ORIS, J. T. (1997). Estimating inhibition concentrations for different response scales using generalized linear models. *Environ. Toxicol. Chem.* **16** 1554–1559.

BAILEY, H. C., DIGIORGIO, C., KROLL, K., HINTON, D. E., MILLER, J. L. and STARRETT, G. (1996). Development of procedures for identifying pesticide toxicity in ambient waters: Carbofuran, diazinon and chlorpyrifos. *Environ. Toxicol. Chem.* **15** 837–845.

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

BURDEN, N., GELLATLY, N., BENSTEAD, R., BENYON, K., BLICKLEY, T. M., CLOOK, M., DOYLE, I., EDWARDS, P., HANDLEY, J. et al. (2017). Reducing repetition of regulatory vertebrate ecotoxicology studies. *Integr. Environ. Assess. Manag.* **13** 955–957. https://doi.org/10.1002/ieam.1934

CAI, D., CAMPBELL, T. and BRODERICK, T. (2021). Finite mixture models do not reliably learn the number of components. In *International Conference on Machine Learning* 1158–1169.

DOBSON, A. J. and BARNETT, A. G. (2018). *An Introduction to Generalized Linear Models*, 4th ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. For the third edition see [MR2459739]. MR3890007

DUDOIT, S. and FRIDLYAND, J. (2021). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.* **3** 1–21.

FAN, Y. and LI, R. (2012). Variable selection in linear mixed effects models. *Ann. Statist.* **40** 2043–2068. MR3059076 https://doi.org/10.1214/12-AOS1028

FANG, K., CHEN, Y., MA, S. and ZHANG, Q. (2022). Biclustering analysis of functionals via penalized fusion. *J. Multivariate Anal.* **189** Paper No. 104874, 20. MR4384116 https://doi.org/10.1016/j.jmva.2021.104874

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635 https://doi.org/10.1198/016214502760047131

Key words and phrases. Aquatic toxicity assessment, heterogeneous variance, linear mixed models, penalty functions.

FRÜHWIRTH-SCHNATTER, S., MALSINER-WALLI, G. and GRÜN, B. (2021). Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Anal.* **16** 1279–1307. MR4381135 https://doi.org/10.1214/21-BA1294

HU, X., HUANG, J., LIU, L., SUN, D. and ZHAO, X. (2021). Subgroup analysis in the heterogeneous Cox model. *Stat. Med.* **40** 739–757. MR4198442 https://doi.org/10.1002/sim.8800

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.

JAIN, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recogn. Lett.* **31** 651–666.

LEISCH, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *J. Stat. Softw.* **11** 1–18. https://doi.org/10.18637/jss.v011.i08

LV, Y., ZHU, X., ZHU, Z. and QU, A. (2020). Nonparametric cluster analysis on multiple outcomes of longitudinal data. *Statist. Sinica* **30** 1829–1856. MR4260746 https://doi.org/10.5705/ss.202018.0032

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

MA, S., HUANG, J., ZHANG, Z. and LIU, M. (2020). Exploration of heterogeneous treatment effects via concave fusion. *Int. J. Biostat.* **16**.

MALSINER-WALLI, G., FRÜHWIRTH-SCHNATTER, S. and GRÜN, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* **26** 303–324. MR3439375 https://doi.org/10.1007/s11222-014-9500-2

MILJKOVIC, T. and WANG, X. (2021). Identifying subgroups of age and cohort effects in obesity prevalence. *Biom. J.* **63** 168–186. MR4204907 https://doi.org/10.1002/bimj.201900287

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

STROUP, W. W. (2013). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. Texts in Statistical Science Series*. CRC Press. MR2977489

TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 411–423. MR1841503 https://doi.org/10.1111/1467-9868.00293

VINH, N. X., EPPS, J. and BAILEY, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11** 2837–2854. MR2738784

WANG, H., LI, R. and TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. MR2410008 https://doi.org/10.1093/biomet/asm053

WANG, X. (2024). Clustering of longitudinal curves via a penalized method and EM algorithm. *Comput. Statist.* **39** 1485–1512. MR4730670

WANG, X. and ZHANG, J. (2024). Supplement to “Assessing aquatic toxicity assessment via a clustered variance model.” https://doi.org/10.1214/24-AOAS1884SUPPA, https://doi.org/10.1214/24-AOAS1884SUPPB

WANG, X., ZHANG, X. and ZHU, Z. (2023). Clustered coefficient regression models for Poisson process with an application to seasonal warranty claim data. *Technometrics* **65** 514–523. MR4662685 https://doi.org/10.1080/00401706.2023.2190779

WANG, X. and ZHU, Z. (2019). Small area estimation with subgroup analysis. *Stat. Theory Relat. Fields* **3** 129–135. MR4028311 https://doi.org/10.1080/24754269.2019.1659097

WANG, X., ZHU, Z. and ZHANG, H. H. (2023). Spatial heterogeneity automatic detection and estimation. *Comput. Statist. Data Anal.* **180** Paper No. 107667, 23. MR4519305 https://doi.org/10.1016/j.csda.2022.107667

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

ZHANG, J., KONG, Y., BAILER, A. J., ZHU, Z. and SMUCKER, B. (2022). Incorporating historical data when determining sample size requirements for aquatic toxicity experiments. *J. Agric. Biol. Environ. Stat.* **27** 544–561. MR4459080 https://doi.org/10.1007/s13253-022-00496-0

ZHOU, L., SUN, S., FU, H. and SONG, P. X.-K. (2022). Subgroup-effects models for the analysis of personal treatment effects. *Ann. Appl. Stat.* **16** 80–103. MR4400504 https://doi.org/10.1214/21-aoas1503

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

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2359–2381
https://doi.org/10.1214/24-AOAS1885

A LATENT VARIABLE APPROACH FOR MODELING RELATIONAL DATA WITH MULTIPLE RECEIVERS

BY JORIS MULDER[1,a] AND PETER D. HOFF[2,b]

[1]*Department of Methodology and Statistics, Tilburg University,* [a]*j.mulder3@tilburguniversity.edu*
[2]*Department of Statistical Science, Duke University,* [b]*peter.hoff@duke.edu*

Directional relational event data, such as email data, often contain unicast messages (i.e., messages of one sender toward one receiver) and multicast messages (i.e., messages of one sender toward multiple receivers). The Enron email data that is the focus in this paper consists of 31% multicast messages. Multicast messages contain important information about the roles of actors in the network, which is needed for better understanding social interaction dynamics. In this paper a multiplicative latent factor model is proposed to analyze such relational data. For a given message, all potential receiver actors are placed on a suitability scale, and the actors are included in the receiver set whose suitability score exceeds a threshold value. Unobserved heterogeneity in the social interaction behavior is captured using a multiplicative latent factor structure with latent variables for actors (which differ for actors as senders and receivers) and latent variables for individual messages. The model is referred to as the multicast additive and multiplicative effects network (mc-amen) model. A Bayesian computational algorithm, which relies on Gibbs sampling, is proposed for model fitting. Model assessment is done using posterior predictive checks. Numerical simulations show that the model is widely applicable for various scenarios involving multicast messages. Furthermore, a mc-amen model with a two-dimensional latent variable can accurately capture the empirical distribution of the cardinality of the receiver set and the composition of the receiver sets for commonly observed messages in the Enron email data. In the Enron network, actors have a comparable (but not identical) role as a sender and as a receiver in the network.

REFERENCES

ARENA, G., MULDER, J. and LEENDERS, R. T. A. (2022). A Bayesian semi-parametric approach for modeling memory decay in dynamic social networks. *Sociol. Methods Res.* 1–51.

BRANDES, U., LERNER, J. and SNIJDERS, T. A. B. (2009). Networks evolving step by step: Statistical analysis of dyadic event data. In 2009 *International Conference on Advances in Social Network Analysis and Mining* 200–205.

BUTTS, C. T. (2008). A relational event framework for social action. *Sociol. Method.* **38** 155–200.

CARTWRIGHT, D. and HARARY, F. (1956). Structural balance: A generalization of Heider's theory. *Psychol. Rev.* **63** 277–293. https://doi.org/10.1037/h0046049

COHEN, W. W. (2009). Enron email dataset.

DUBOIS, C., BUTTS, C. and SMYTH, P. (2013). Stochastic blockmodeling of relational event dynamics. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* (C. M. Carvalho and P. Ravikumar, eds.). *Proceedings of Machine Learning Research* **31** 238–246. PMLR, Scottsdale, AZ, USA.

DUBOIS, C., BUTTS, C. T., MCFARLAND, D. and SMYTH, P. (2013). Hierarchical models for relational event sequences. *J. Math. Psych.* **57** 297–309. MR3137883 https://doi.org/10.1016/j.jmp.2013.04.001

ECKMANN, J.-P., MOSES, E. and SERGI, D. (2004). Entropy of dialogues creates coherent structures in e-mail traffic. *Proc. Natl. Acad. Sci. USA* **101** 14333–14337. MR2098979 https://doi.org/10.1073/pnas.0405728101

GABRY, J. and MAHR, T. (2017). bayesplot: Plotting for Bayesian models. R package version 1.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR2027492

Key words and phrases. Relational data, multicast messages, latent variable modeling.

HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354. MR2364300 https://doi.org/10.1111/j.1467-985X.2007.00471.x

HEIDER, F. (1946). Attitudes and cognitive organization. *J. Psychol.* **21** 107–112. https://doi.org/10.1080/00223980.1946.9917275

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. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics* (J. C. Platt, D. Koller, Y. Singer and S. Roweis, eds.). *Advances in Neural Information Processing Systems* **20** 657–664. MIT Press, Cambridge.

HOFF, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Comput. Math. Organ. Theory* **15** 261–272.

HOFF, P. D. (2015). Dyadic data analysis with amen. arXiv preprint. Available at arXiv:1506.08237.

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

LEENDERS, R. T. A. J., CONTRACTOR, N. S. and DECHURCH, L. A. (2016). Once upon a time: Understanding team processes as relational event networks. *Organ. Psychol. Rev.* **6** 92–115. https://doi.org/10.1177/2041386615578312

MEIJERINK-BOSMAN, M., BACK, M., GEUKES, K., LEENDERS, R. and MULDER, J. (2022). Discovering trends of social interaction behavior over time: An introduction to relational event modeling: Trends of social interaction. *Behav. Res. Methods* 1–27.

MEIJERINK-BOSMAN, M., LEENDERS, R. and MULDER, J. (2022). Dynamic relational event modeling: Testing, exploring, and applying. *PLoS ONE* **17** e0272309. https://doi.org/10.1371/journal.pone.0272309

MENG, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.* **22** 1142–1160. MR1311969 https://doi.org/10.1214/aos/1176325622

MULDER, J. and HOFF, P. D. (2024). Supplement to “A latent variable approach for modeling relational data with multiple receivers.” https://doi.org/10.1214/24-AOAS1885SUPP

MULDER, J. and LEENDERS, R. T. A. J. (2019). Modeling the evolution of interaction behavior in social networks: A dynamic relational event approach for real-time analysis. *Chaos Solitons Fractals* **119** 73–85. MR3894308 https://doi.org/10.1016/j.chaos.2018.11.027

NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96** 1077–1087. MR1947255 https://doi.org/10.1198/016214501753208735

PERRY, P. O. and WOLFE, P. J. (2013). Point process modelling for directed interaction networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 821–849. MR3124793 https://doi.org/10.1111/rssb.12013

QUINTANE, E., CONALDI, G., TONELLATO, M. and LOMI, A. (2014). Modeling relational events: A case study on an open source software project. *Organ. Res. Methods* **17** 23–50. https://doi.org/10.1177/1094428113517007

SHAFIEI, M. and CHIPMAN, H. (2010). Mixed-membership stochastic block-models for transactional networks. In 2010 *IEEE International Conference on Data Mining*. https://doi.org/10.1109/ICDM.2010.88.

STADTFELD, C. and BLOCK, P. (2017). Interactions, actors, and time: Dynamic network actor models for relational events. *Sociol. Sci.* **4** 318–352. https://doi.org/10.15195/v4.a14

VAN KOLLENBURG, G. H., MULDER, J. and VERMUNT, J. K. (2015). Assessing model fit in latent class analysis when asymptotics do not hold. *Methodology* **1** 65–79.

ZHOU, Y., GOLDBERG, M., MAGDON-ISMAIL, M. and WALLACE, A. (2007). Strategies for cleaning organizational emails with an application to enron email dataset. In 5*th Conf. of North American Association for Computational Social and Organizational Science* **0621303**.

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2382–2402
https://doi.org/10.1214/24-AOAS1886

SPARSE CLUSTERING FOR CUSTOMER SEGMENTATION WITH HIGH-DIMENSIONAL MIXED-TYPE DATA

BY FEIFEI WANG[1,a], SHAODONG XU[1,b], YICHEN QIN[2,d], YE SHEN[3,e] AND YANG LI[1,c]

[1]*Center for Applied Statistics and School of Statistics, Renmin University of China,* [a]*feifei.wang@ruc.edu.cn,* [b]*shaodong_xu@ruc.edu.cn,* [c]*yang.li@ruc.edu.cn*

[2]*Department of Operations, Business Analytics and Information Systems, University of Cincinnati,* [d]*qinyn@ucmail.uc.edu*

[3]*Department of Epidemiology and Biostatistics, University of Georgia,* [e]*yeshen@uga.edu*

Customer segmentation has wide applications in business activities, such as personalized marketing and targeted product development. To realize customer segmentation, clustering methods are commonly used. However, modern customer segmentation encounters challenges characterized by high-dimensionality and mixed-type variables (i.e., the mixture of continuous variables and categorical variables). It brings great challenges to customer segmentation, because most existing clustering methods are only designed for data with one single type of variables. Furthermore, the existence of noise variables highlights the necessity of simultaneous variable selection and data clustering. Motivated by these issues, we develop a Davies–Bouldin index based sparse clustering (DBI-SC) method for customer segmentation with high-dimensional mixed-type data. In this method we define dissimilarity measures for continuous variables and categorical variables separately. Then an adjusted DBI criterion is designed to measure the contribution of each variable to clustering. For variable selection we apply the sparse clustering framework and introduce different penalty parameters for the mixed-type variables. The screening consistency property of the DBI-SC method is also investigated. Extensive simulation studies demonstrate the satisfactory performance of the DBI-SC method in both clustering and variable selection. Finally, a designated driving service dataset is analyzed for customer segmentation using the proposed method.

REFERENCES

AHMAD, A. and DEY, L. (2007). A K-means clustering algorithm for mixed numeric and categorical data. *Data Knowl. Eng.* **63** 503–527.

ALELYANI, S., TANG, J. and LIU, H. (2018). Feature selection for clustering: A review. *Data Clustering* 29–60.

ARIAS-CASTRO, E. and PU, X. (2017). A simple approach to sparse clustering. *Comput. Statist. Data Anal.* **105** 217–228. MR3552198 https://doi.org/10.1016/j.csda.2016.08.003

BALLESTAR, M. T., GRAU-CARLES, P. and SAINZ, J. (2018). Customer segmentation in E-commerce: Applications to the cashback business model. *J. Bus. Res.* **88** 407–414.

CHAVENT, M., KUENTZ-SIMONET, V., LABENNE, A. and SARACCO, J. (2014). Multivariate analysis of mixed data: The R package PCAmixdata. ArXiv preprint. Available at arXiv:1411.4911.

CHAVENT, M., LACAILLE, J., MOURER, A. and OLTEANU, M. (2020). Sparse K-means for mixed data via group-sparse clustering. In *ESANN* 2020-28*th European Symposium on Artificial Neural Networks*, *Computational Intelligence and Machine Learning* **978**.

CHUNG, J., JOO, H. H. and MOON, S. (2014). Designated driver service availability and its effects on drunk driving behaviors. *B.E. J. Econ. Anal. Policy* **14** 1543–1567.

DAVIES, D. L. and BOULDIN, D. W. (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **2** 224–227.

DORMAN, K. S. and MAITRA, R. (2022). An efficient k-modes algorithm for clustering categorical datasets. *Stat. Anal. Data Min.* **15** 83–97. MR4400768 https://doi.org/10.1002/sam.11546

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

Key words and phrases. Customer segmentation, Davies–Bouldin Index (DBI), heterogeneity analysis, unsupervised learning, variable selection.

FOP, M., SMART, K. M. and MURPHY, T. B. (2017). Variable selection for latent class analysis with application to low back pain diagnosis. *Ann. Appl. Stat.* **11** 2080–2110. MR3743289 https://doi.org/10.1214/17-AOAS1061

FOSS, A., MARKATOU, M., RAY, B. and HECHING, A. (2016). A semiparametric method for clustering mixed data. *Mach. Learn.* **105** 419–458. MR3557450 https://doi.org/10.1007/s10994-016-5575-7

FU, Y., LIU, X., SARKAR, S. and WU, T. (2021). Gaussian mixture model with feature selection: An embedded approach. *Comput. Ind. Eng.* **152** 107000.

GUTTENTAG, D., SMITH, S., POTWARKA, L. and HAVITZ, M. (2018). Why tourists choose airbnb: A motivation-based segmentation study. *J. Travel Res.* **57** 342–359.

HUANG, J. and MA, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Anal.* **16** 176–195. MR2608284 https://doi.org/10.1007/s10985-009-9144-2

HUANG, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the* 1*st Pacific-Asia Conference on Knowledge Discovery and Data Mining*, (*PAKDD*) 21–34. Citeseer.

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.

INSIGHT and INFO (2022). In Depth Research on the Status Quo of China's Designated Driving Industry and Future Investment Forecast Report (2022-2029). Available at https://www.chinabaogao.com/baogao/202209/608779.html.

JOE, H. (2006). Generating random correlation matrices based on partial correlations. *J. Multivariate Anal.* **97** 2177–2189. MR2301633 https://doi.org/10.1016/j.jmva.2005.05.010

JOU, R. and SYU, L. (2021). Drunk drivers' willingness to use and to pay for designated drivers. *Sustainability* **13** 5362.

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

LAGONA, F. and PICONE, M. (2022). A latent-class model for clustering incomplete linear and circular data in marine studies. *J. Data Sci.* **9** 585–605.

LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274

MARBAC, M. and SEDKI, M. (2017). Variable selection for mixed data clustering: A model-based approach. ArXiv preprint. Available at arXiv:1703.02293.

MARBAC, M., SEDKI, M. and PATIN, T. (2020). Variable selection for mixed data clustering: Application in human population genomics. *J. Classification* **37** 124–142. MR4111887 https://doi.org/10.1007/s00357-018-9301-y

MCPARLAND, D. and GORMLEY, I. C. (2016). Model based clustering for mixed data: ClustMD. *Adv. Data Anal. Classif.* **10** 155–169. MR3505054 https://doi.org/10.1007/s11634-016-0238-x

NAKANO, S. and KONDO, F. N. (2018). Customer segmentation with purchase channels and media touchpoints using single source panel data. *J. Retail. Consum. Serv.* **41** 142–152.

SCHUBERT, E. and ROUSSEEUW, P. J. (2021). Fast and eager K-medoids clustering: O (k) runtime improvement of the PAM, Clara, and CLARANS algorithms. *Inform. Sci.* **101** 101804.

SILVESTRE, C., CARDOSO, M. G. and FIGUEIREDO, M. (2015). Feature selection for clustering categorical data with an embedded modelling approach. *Expert Syst.* **32** 444–453.

STORLIE, C. B., MYERS, S. M., KATUSIC, S. K., WEAVER, A. L., VOIGT, R. G., CROARKIN, P. E., STOECKEL, R. E. and PORT, J. D. (2018). Clustering and variable selection in the presence of mixed variable types and missing data. *Stat. Med.* **37** 2884–2899. MR3832247 https://doi.org/10.1002/sim.7697

WANG, F., XU, S., QIN, Y., SHEN, Y. and LI, Y. (2024). Supplement to "Sparse clustering for customer segmentation with high-dimensional mixed-type data." https://doi.org/10.1214/24-AOAS1886SUPPA, https://doi.org/10.1214/24-AOAS1886SUPPB

WITTEN, D. M. and TIBSHIRANI, R. (2010). A framework for feature selection in clustering. *J. Amer. Statist. Assoc.* **105** 713–726. MR2724855 https://doi.org/10.1198/jasa.2010.tm09415

YE, M., ZHANG, P. and NIE, L. (2018). Clustering sparse binary data with hierarchical Bayesian Bernoulli mixture model. *Comput. Statist. Data Anal.* **123** 32–49. MR3777084 https://doi.org/10.1016/j.csda.2018.01.020

ZHOU, Z. and AMINI, A. A. (2019). Analysis of spectral clustering algorithms for community detection: The general bipartite setting. *J. Mach. Learn. Res.* **20** 47. MR3948087

ZHU, Y., DENG, Q., HUANG, D., JING, B. and ZHANG, B. (2021). Clustering based on Kolmogorov-Smirnov statistic with application to bank card transaction data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **70** 558–578.

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2403–2423
https://doi.org/10.1214/24-AOAS1887

EVALUATION OF TRANSPLANT BENEFITS WITH THE U.S. SCIENTIFIC REGISTRY OF TRANSPLANT RECIPIENTS BY SEMIPARAMETRIC REGRESSION OF MEAN RESIDUAL LIFE

BY GE ZHAO[1,a], YANYUAN MA[2,b], HUAZHEN LIN[3,c] AND YI LI[4,d]

[1]*Department of Mathematics and Statistics, Portland State University,* [a]*gzhao@pdx.edu*
[2]*Department of Statistics, Pennsylvania State University,* [b]*yzm63@psu.edu*
[3]*Center of Statistical Research, Southwestern University of Finance and Economics,* [c]*linhz@swufe.edu.cn*
[4]*Department of Biostatistics, University of Michigan, Ann Arbor,* [d]*yili@umich.edu*

Kidney transplantation is the most effective renal replacement therapy for end stage renal disease patients. With the severe shortage of kidney supplies and for the clinical effectiveness of transplantation, patient's life expectancy posttransplantation is used to prioritize patients for transplantation; however, severe comorbidity conditions and old age are the most dominant factors that negatively impact posttransplantation life expectancy, effectively precluding sick or old patients from receiving transplants. It would be crucial to design objective measures to quantify the transplantation benefit by comparing the mean residual life with and without a transplant, after adjusting for comorbidity and demographic conditions. To address this urgent need, we propose a new class of semiparametric covariate-dependent mean residual life models. Our method estimates covariate effects semiparametrically efficiently and the mean residual life function nonparametrically, enabling us to predict the residual life increment potential for any given patient. Our method potentially leads to a more fair system that prioritizes patients who would have the largest residual life gains. Our analysis of the kidney transplant data from the U.S. Scientific Registry of Transplant Recipients also suggests that a single index of covariates summarize well the impacts of multiple covariates, which may facilitate interpretations of each covariate's effect. Our subgroup analysis further disclosed inequalities in survival gains across groups defined by race, gender and insurance type (reflecting socioeconomic status).

REFERENCES

AALEN, O. O., COOK, R. J. and RØYSLAND, K. (2015). Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Anal.* **21** 579–593. MR3397507 https://doi.org/10.1007/s10985-015-9335-y

AHMADI, S.-F., ZAHMATKESH, G., AHMADI, E., STREJA, E., RHEE, C. M., GILLEN, D. L., DE NICOLA, L., MINUTOLO, R., RICARDO, A. C. et al. (2016). Association of body mass index with clinical outcomes in non-dialysis-dependent chronic kidney disease: A systematic review and meta-analysis. *Cardiorenal Med.* **6** 37–49.

ALI, H., SOLIMAN, K., MOHAMED, M. M., RAHMAN, M., HERBERTH, J., FÜLÖP, T. and ELSAYED, I. (2021). Impact of kidney transplantation on functional status. *Ann. Med.* **53** 1303–1309.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. *Springer Series in Statistics*. Springer, New York. MR1198884 https://doi.org/10.1007/978-1-4612-4348-9

ANDERSEN, P. K., SYRIOPOULOU, E. and PARNER, E. T. (2017). Causal inference in survival analysis using pseudo-observations. *Stat. Med.* **36** 2669–2681. MR3670384 https://doi.org/10.1002/sim.7297

ASSFALG, V., SELIG, K., TOLKSDORF, J., VAN MEEL, M., DE VRIES, E., RAMSOEBHAG, A.-M., RAHMEL, A., RENDERS, L., NOVOTNY, A. et al. (2020). Repeated kidney re-transplantation—the Eurotransplant experience: A retrospective multicenter outcome analysis. *Transpl. Int.* **33** 617–631.

Key words and phrases. Mean residual life, kidney transplant, semiparametric efficient estimator, nonparametric estimation.

AXELROD, D. A., GUIDINGER, M. K., FINLAYSON, S., SCHAUBEL, D. E., GOODMAN, D. C., CHOBANIAN, M. and MERION, R. M. (2008). Rates of solid-organ wait-listing, transplantation, and survival among residents of rural and urban areas. *JAMA* **299** 202–207.

BAYLIS, C. (2009). Sexual dimorphism in the aging kidney: Differences in the nitric oxide system. *Nat. Rev. Nephrol.* **5** 384–396. https://doi.org/10.1038/nrneph.2009.90

BEDDHU, S. (2004). Hypothesis: The body mass index paradox and an obesity, inflammation, and atherosclerosis syndrome in chronic kidney disease. In *Seminars in Dialysis* **17** 229–232. Wiley, New York.

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

BUI, K., KILAMBI, V. and MEHROTRA, S. (2019). Functional status-based risk-benefit analyses of high-KDPI kidney transplant versus dialysis. *Transplant. Int.* **32** 1297–1312. https://doi.org/10.1111/tri.13483

CARRERO, J. J. (2010). Gender differences in chronic kidney disease: Underpinnings and therapeutic implications. *Kidney Blood Press. Res.* **33** 383–392. https://doi.org/10.1159/000320389

CARRERO, J. J., HECKING, M., CHESNAYE, N. C. and JAGER, K. J. (2018). Sex and gender disparities in the epidemiology and outcomes of chronic kidney disease. *Nat. Rev. Nephrol.* **14** 151–164. https://doi.org/10.1038/nrneph.2017.181

CASELLA, G. and BERGER, R. L. (2001). *Statistical Inference*. Cengage Learning. MR1051420

CHADBAN, S. J., AHN, C., AXELROD, D. A., FOSTER, B. J., KASISKE, B. L., KHER, V., KUMAR, D., OBERBAUER, R., PASCUAL, J. et al. (2020). KDIGO clinical practice guideline on the evaluation and management of candidates for kidney transplantation. *Transplantation* **104** S11–S103.

CHEN, Y. Q. (2007). Additive expectancy regression. *J. Amer. Statist. Assoc.* **102** 153–166. MR2345536 https://doi.org/10.1198/016214506000000870

CHEN, Y. Q. and CHENG, S. (2005). Semiparametric regression analysis of mean residual life with censored survival data. *Biometrika* **92** 19–29. MR2158607 https://doi.org/10.1093/biomet/92.1.19

CHEN, Y. Q. and CHENG, S. (2006). Linear life expectancy regression with censored data. *Biometrika* **93** 303–313. MR2278085 https://doi.org/10.1093/biomet/93.2.303

CHEN, Y. Q., JEWELL, N. P., LEI, X. and CHENG, S. C. (2005). Semiparametric estimation of proportional mean residual life model in presence of censoring. *Biometrics* **61** 170–178. MR2135857 https://doi.org/10.1111/j.0006-341X.2005.030224.x

COBO, G., HECKING, M., PORT, F. K., EXNER, I., LINDHOLM, B., STENVINKEL, P. and CARRERO, J. J. (2016). Sex and gender differences in chronic kidney disease: Progression to end-stage renal disease and haemodialysis. *Clin. Sci.* **130** 1147–1163. https://doi.org/10.1042/CS20160047

COSIO, F. G., ALAMIR, A., YIM, S., PESAVENTO, T. E., FALKENHAIN, M. E., HENRY, M. L., ELKHAMMAS, E. A., DAVIES, E. A., BUMGARDNER, G. L. et al. (1998). Patient survival after renal transplantation: I. The impact of dialysis pre-transplant. *Kidney Inter.* **53** 767–772. https://doi.org/10.1046/j.1523-1755.1998.00787.x

DEN DEKKER, W. K., SLOT, M. C., KHO, M. M. L., GALEMA, T. W., VAN DE WETERING, J., BOERSMA, E. and ROODNAT, J. I. (2020). Predictors of postoperative cardiovascular complications up to 3 months after kidney transplantation. *Neth. Heart J.* **28** 202–209. https://doi.org/10.1007/s12471-020-01373-6

EVANS, R. W., MANNINEN, D. L., GARRISON, L. P. JR, HART, L. G., BLAGG, C. R., GUTMAN, R. A., HULL, A. R. and LOWRIE, E. G. (1985). The quality of life of patients with end-stage renal disease. *N. Engl. J. Med.* **312** 553–559.

FEDEWA, S. A., MCCLELLAN, W. M., JUDD, S., GUTIÉRREZ, O. M. and CREWS, D. C. (2014). The association between race and income on risk of mortality in patients with moderate chronic kidney disease. *BMC Nephrol.* **15** 1–9.

FENG, Y., HUANG, R., KAVANAGH, J., LI, L., ZENG, X., LI, Y. and FU, P. (2019). Efficacy and safety of dual blockade of the renin–angiotensin–aldosterone system in diabetic kidney disease: A meta-analysis. *Amer. J. Cardiovasc. Drugs* **19** 259–286.

FERRI, F. F. (2017). *Ferri's Clinical Advisor* 2018 *e-Book*: 5 *Books in* 1. Elsevier, MO.

FRIEDMAN, A. N., MISKULIN, D. C., ROSENBERG, I. H. and LEVEY, A. S. (2003). Demographics and trends in overweight and obesity in patients at time of kidney transplantation. *Amer. J. Kidney Dis.* **41** 480–487. https://doi.org/10.1053/ajkd.2003.50059

GOLDFARB-RUMYANTZEV, A. S., KOFORD, J. K., BAIRD, B. C., CHELAMCHARLA, M., HABIB, A. N., WANG, B.-J., LIN, S., SHIHAB, F. and ISAACS, R. B. (2006). Role of socioeconomic status in kidney transplant outcome. *Clin. J. Amer. Soc. Nephrol.* **1** 313–322. https://doi.org/10.2215/CJN.00630805

GORE, J. L., DANOVITCH, G. M., LITWIN, M. S., PHAM, P. T. and SINGER, J. S. (2009). Disparities in the utilization of live donor renal transplantation. *Amer. J. Transplant.* **9** 1124–1133. https://doi.org/10.1111/j.1600-6143.2009.02620.x

HALL, W. J. and WELLNER, J. (1981). Mean residual life. In *Statistics and Related Topics* (*Ottawa*, *Ont*., 1980) 169–184. North-Holland, Amsterdam. MR0665274

HART, A., LENTINE, K., SMITH, J., MILLER, J., SKEANS, M., PRENTICE, M., ROBINSON, A., FOUTZ, J., BOOKER, S. et al. (2021). OPTN/SRTR 2019 annual data report: Kidney. *Amer. J. Transplant*. **21** 21–137.

HERNANDEZ, D., RUFINO, M., ARMAS, S., GONZALEZ, A., GUTIERREZ, P., BARBERO, P., VIVANCOS, S., RODRÍGUEZ, C., DE VERA, J. R. et al. (2006). Retrospective analysis of surgical complications following cadaveric kidney transplantation in the modern transplant era. *Nephrol. Dial. Transplant*. **21** 2908–2915.

HUMAR, A. and MATAS, A. J. (2005). Surgical complications after kidney transplantation. In *Seminars in Dialysis* **18** 505–510. Wiley, New York.

IIDA, S., KONDO, T., AMANO, H., NAKAZAWA, H., ITO, F., HASHIMOTO, Y. and TANABE, K. (2008). Minimal effect of cold ischemia time on progression to late-stage chronic kidney disease observed long term after partial nephrectomy. *Urology* **72** 1083–1088.

INGSATHIT, A., KAMANAMOOL, N., THAKKINSTIAN, A. and SUMETHKUL, V. (2013). Survival advantage of kidney transplantation over dialysis in patients with hepatitis C: A systematic review and meta-analysis. *Transplantation* **95** 943–948. https://doi.org/10.1097/TP.0b013e3182848de2

JASSAL, S. V., SCHAUBEL, D. E. and FENTON, S. S. A. (2005). Baseline comorbidity in kidney transplant recipients: A comparison of comorbidity indices. *Amer. J. Kidney Dis*. **46** 136–142. https://doi.org/10.1053/j.ajkd.2005.03.006

JEONG, J.-H., JUNG, S.-H. and COSTANTINO, J. P. (2008). Nonparametric inference on median residual life function. *Biometrics* **64** 157–163, 323–324. MR2422830 https://doi.org/10.1111/j.1541-0420.2007.00826.x

JUNG, S.-H., JEONG, J.-H. and BANDOS, H. (2009). Regression on quantile residual life. *Biometrics* **65** 1203–1212. MR2756508 https://doi.org/10.1111/j.1541-0420.2009.01196.x

KALANTAR-ZADEH, K., ABBOTT, K. C., SALAHUDEEN, A. K., KILPATRICK, R. D. and HORWICH, T. B. (2005). Survival advantages of obesity in dialysis patients. *Amer. J. Clin. Nutr.* **81** 543–554.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0570114

KASISKE, B. L., CANGRO, C. B., HARIHARAN, S., HRICIK, D. E., KERMAN, R. H., ROTH, D., RUSH, D. N., VAZQUEZ, M. A., WEIR, M. R. et al. (2001). The evaluation of renal transplantation candidates: Clinical practice guidelines. *Amer. J. Transplant*. **1** 3–95.

KASISKE, B. L., LONDON, W. and ELLISON, M. D. (1998). Race and socioeconomic factors influencing early placement on the kidney transplant waiting list. *J. Amer. Soc. Nephrol*. **9** 2142–2147. https://doi.org/10.1681/ASN.V9112142

KAUFFMAN, H. M., CHERIKH, W. S., MCBRIDE, M. A., CHENG, Y. A., DELMONICO, F. L. and HANTO, D. W. (2005). Transplant recipients with a history of a malignancy: Risk of recurrent and de novo cancers. *Transplant. Rev.* **19** 55–64.

KAYLER, L. K., MAGLIOCCA, J., ZENDEJAS, I., SRINIVAS, T. R. and SCHOLD, J. D. (2011). Impact of cold ischemia time on graft survival among ECD transplant recipients: A paired kidney analysis. *Amer. J. Transplant*. **11** 2647–2656. https://doi.org/10.1111/j.1600-6143.2011.03741.x

KUCIRKA, L. M., PURNELL, T. S. and SEGEV, D. L. (2015). Improving access to kidney transplantation: Referral is not enough. *JAMA* **314** 565–567.

LEWIS, J. et al. (2010). Racial differences in chronic kidney disease (CKD) and end-stage renal disease (ESRD) in the United States: A social and economic dilemma. *Clin. Nephrol*. **74** S72–7.

LIEM, Y. S., BOSCH, J. L., ARENDS, L. R., HEIJENBROK-KAL, M. H. and HUNINK, M. M. (2007). Quality of life assessed with the medical outcomes study short form 36-item health survey of patients on renal replacement therapy: A systematic review and meta-analysis. *Value Health* **10** 390–397.

LIEM, Y. S. and WEIMAR, W. (2009). Early living-donor kidney transplantation: A review of the associated survival benefit. *Transplantation* **87** 317–318. https://doi.org/10.1097/TP.0b013e3181952710

LIM, W. H., CHAPMAN, J. R. and WONG, G. (2015). Peak panel reactive antibody, cancer, graft, and patient outcomes in kidney transplant recipients. *Transplantation* **99** 1043–1050.

LIN, H., FEI, Z. and LI, Y. (2016). A semiparametrically efficient estimator of the time-varying effects for survival data with time-dependent treatment. *Scand. J. Stat*. **43** 649–663. MR3543315 https://doi.org/10.1111/sjos.12196

MA, Y. and WEI, Y. (2012). Analysis on censored quantile residual life model via spline smoothing. *Statist. Sinica* **22** 47–68. MR2933167 https://doi.org/10.5705/ss.2010.161

MA, Y. and YIN, G. (2010). Semiparametric median residual life model and inference. *Canad. J. Statist*. **38** 665–679. MR2753008 https://doi.org/10.1002/cjs.10076

MA, Y. and ZHANG, X. (2015). A validated information criterion to determine the structural dimension in dimension reduction models. *Biometrika* **102** 409–420. MR3371013 https://doi.org/10.1093/biomet/asv004

MA, Y. and ZHU, L. (2012). A semiparametric approach to dimension reduction. *J. Amer. Statist. Assoc*. **107** 168–179. MR2949349 https://doi.org/10.1080/01621459.2011.646925

MA, Y. and ZHU, L. (2013). Efficient estimation in sufficient dimension reduction. *Ann. Statist.* **41** 250–268. MR3059417 https://doi.org/10.1214/12-AOS1072

MAFRA, D., GUEBRE-EGZIABHER, F. and FOUQUE, D. (2008). Body mass index, muscle and fat in chronic kidney disease: Questions about survival. *Nephrol. Dial. Transplant.* **23** 2461–2466.

MAGULURI, G. and ZHANG, C.-H. (1994). Estimation in the mean residual life regression model. *J. Roy. Statist. Soc. Ser. B* **56** 477–489. MR1278221

MANSOURVAR, Z., MARTINUSSEN, T. and SCHEIKE, T. H. (2015). Semiparametric regression for restricted mean residual life under right censoring. *J. Appl. Stat.* **42** 2597–2613. MR3428833 https://doi.org/10.1080/02664763.2015.1043871

MANSOURVAR, Z., MARTINUSSEN, T. and SCHEIKE, T. H. (2016). An additive-multiplicative restricted mean residual life model. *Scand. J. Stat.* **43** 487–504. MR3503013 https://doi.org/10.1111/sjos.12187

MEHDI, U. and TOTO, R. D. (2009). Anemia, diabetes, and chronic kidney disease. *Diabetes Care* **32** 1320–1326. https://doi.org/10.2337/dc08-0779

MEIER-KRIESCHE, H.-U., PORT, F. K., OJO, O. A., AKINLOLU, O., RUDICH, S. M., HANSON, J. A., CIBRIK, D. M., LEICHTMAN, A. B. and KAPLAN, B. (2000). Effect of waiting time on renal transplant outcome. *Kidney Inter.* **58** 1311–1317.

MOLNAR, M. Z., CZIRA, M. E., RUDAS, A., UJSZASZI, A., HAROMSZEKI, B., KOSA, J. P., LAKATOS, P., BEKO, G., SARVARY, E. et al. (2011). Association between the malnutrition-inflammation score and post-transplant anaemia. *Nephrol. Dial. Transplant.* **26** 2000–2006.

MÜLLER, H.-G. and ZHANG, Y. (2005). Time-varying functional regression for predicting remaining lifetime distributions from longitudinal trajectories. *Biometrics* **61** 1064–1075. MR2216200 https://doi.org/10.1111/j.1541-0420.2005.00378.x

MUNTNER, P., NEWSOME, B., KRAMER, H., PERALTA, C. A., KIM, Y., JACOBS, D. R., KIEFE, C. I. and LEWIS, C. E. (2012). Racial differences in the incidence of chronic kidney disease. *Clin. J. Amer. Soc. Nephrol.* **7** 101–107. https://doi.org/10.2215/CJN.06450611

NICHOLAS, S. B., KALANTAR-ZADEH, K. and NORRIS, K. C. (2013). Racial disparities in kidney disease outcomes. In *Seminars in Nephrology* **33** 409–415. Elsevier, Amsterdam.

NICHOLAS, S. B., KALANTAR-ZADEH, K. and NORRIS, K. C. (2015). Socioeconomic disparities in chronic kidney disease. *Adv. Chronic Kidney Dis.* **22** 6–15. https://doi.org/10.1053/j.ackd.2014.07.002

OAKES, D. and DASU, T. (1990). A note on residual life. *Biometrika* **77** 409–410. MR1064816 https://doi.org/10.1093/biomet/77.2.409

OAKES, D. and DASU, T. (2003). Inference for the proportional mean residual life model. In *Crossing Boundaries: Statistical Essays in Honor of Jack Hall. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **43** 105–116. IMS, Beachwood, OH. MR2125050 https://doi.org/10.1214/lnms/1215092393

ØIEN, C. M., REISÆTER, A. V., OS, I., JARDINE, A., FELLSTRÖM, B. and HOLDAAS, H. (2006). Gender-associated risk factors for cardiac end points and total mortality after renal transplantation: Post hoc analysis of the ALERT study. *Clin. Transplant.* **20** 374–382.

OKADA, K., YANAI, M., TAKEUCHI, K., MATSUYAMA, K., NITTA, K., HAYASHI, K. and TAKAHASHI, S. (2014). Sex differences in the prevalence, progression, and improvement of chronic kidney disease. *Kidney Blood Press. Res.* **39** 279–288. https://doi.org/10.1159/000355805

PSCHEIDT, C., NAGEL, G., ZITT, E., KRAMAR, R., CONCIN, H. and LHOTTA, K. (2015). Sex- and time-dependent patterns in risk factors of end-stage renal disease: A large Austrian cohort with up to 20 years of follow-up. *PLoS ONE* **10** e0135052. https://doi.org/10.1371/journal.pone.0135052

PYRAM, R., KANSARA, A., BANERJI, M. A. and LONEY-HUTCHINSON, L. (2012). Chronic kidney disease and diabetes. *Maturitas* **71** 94–103. https://doi.org/10.1016/j.maturitas.2011.11.009

RAHNEMAI-AZAR, A. A., GILCHRIST, B. F. and KAYLER, L. K. (2015). Independent risk factors for early urologic complications after kidney transplantation. *Clin Transplant.* **29** 403–408. https://doi.org/10.1111/ctr.12530

RAMLAU-HANSEN, H. (1983). The choice of a kernel function in the graduation of counting process intensities. *Scand. Actuar. J.* **3** 165–182. MR0724596 https://doi.org/10.1080/03461238.1983.10408700

SALERNO, S., MESSANA, J. M., GREMEL, G. W., DAHLERUS, C., HIRTH, R. A., HAN, P., SEGAL, J. H., XU, T., SHAFFER, D. et al. (2021). COVID-19 risk factors and mortality outcomes among medicare patients receiving long-term dialysis. *JAMA Netw. Open* **4** e2135379–e2135379.

SARAN, R., LI, Y., ROBINSON, B., ABBOTT, K. C., AGODOA, L. Y., AYANIAN, J., BRAGG-GRESHAM, J., BALKRISHNAN, R., CHEN, J. L. et al. (2016). US renal data system 2015 annual data report: Epidemiology of kidney disease in the United States. *Amer. J. Kidney Dis.* **67**.

SCHOLD, J. D., BUCCINI, L. D., GOLDFARB, D. A., FLECHNER, S. M., POGGIO, E. D. and SEHGAL, A. R. (2014). Association between kidney transplant center performance and the survival benefit of transplantation versus dialysis. *Clin. J. Amer. Soc. Nephrol.* **9** 1773–1780. https://doi.org/10.2215/CJN.02380314

SUN, L., SONG, X. and ZHANG, Z. (2012). Mean residual life models with time-dependent coefficients under right censoring. *Biometrika* **99** 185–197. MR2899672 https://doi.org/10.1093/biomet/asr065

SUN, L. and ZHANG, Z. (2009). A class of transformed mean residual life models with censored survival data. *J. Amer. Statist. Assoc.* **104** 803–815. MR2541596 https://doi.org/10.1198/jasa.2009.0130

SYRIOPOULOU, E., RUTHERFORD, M. J. and LAMBERT, P. C. (2020). Marginal measures and causal effects using the relative survival framework. *Int. J. Epidemiol.* **49** 619–628. https://doi.org/10.1093/ije/dyz268

TIME, P. S. (2012). A guide to calculating and interpreting the estimated post-transplant survival (EPTS) score used in the Kidney Allocation System (KAS). Kidney 2.

TONELLI, M., WIEBE, N., KNOLL, G., BELLO, A., BROWNE, S., JADHAV, D., KLARENBACH, S. and GILL, J. (2011). Systematic review: Kidney transplantation compared with dialysis in clinically relevant outcomes. *Amer. J. Transplant.* **11** 2093–2109. https://doi.org/10.1111/j.1600-6143.2011.03686.x

TSIATIS, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18** 354–372. MR1041397 https://doi.org/10.1214/aos/1176347504

WAND, M. P. (1994). Fast computation of multivariate kernel estimators. *J. Comput. Graph. Statist.* **3** 433–445. MR1323051 https://doi.org/10.2307/1390904

WEBSTER, A. C., NAGLER, E. V., MORTON, R. L. and MASSON, P. (2017). Chronic kidney disease. *Lancet* **389** 1238–1252.

WENG, F. L., REESE, P. P., MULGAONKAR, S. and PATEL, A. M. (2010). Barriers to living donor kidney transplantation among black or older transplant candidates. *Clin. J. Amer. Soc. Nephrol.* **5** 2338–2347. https://doi.org/10.2215/CJN.03040410

WOLFE, R. A., ASHBY, V. B., MILFORD, E. L., OJO, A. O., ETTENGER, R. E., AGODOA, L. Y., HELD, P. J. and PORT, F. K. (1999). Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *N. Engl. J. Med.* **341** 1725–1730. https://doi.org/10.1056/NEJM199912023412303

YING, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21** 76–99. MR1212167 https://doi.org/10.1214/aos/1176349016

ZHAO, G., MA, Y., LIN, H. and LI, Y. (2024). Supplement to "Evaluation of transplant benefits with the U.S. Scientific Registry of Transplant Recipients by semiparametric regression of mean residual life." https://doi.org/10.1214/24-AOAS1887SUPP

ZHAO, G., MA, Y. and LU, W. (2022). Efficient estimation for dimension reduction with censored survival data. *Statist. Sinica* **32** 2359–2380. MR4485087 https://doi.org/10.5705/ss.202020.0404

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2424–2443
https://doi.org/10.1214/24-AOAS1888
© Institute of Mathematical Statistics, 2024

SITE OCCUPANCY AND ABUNDANCE MODELS FOR ANALYZING MULTIPLE-VISIT DETECTION/NONDETECTION DATA

BY HUU-DINH HUYNH[1,a], MATTHEW SCHOFIELD[2,b] AND WEN-HAN HWANG[3,c]

[1] *Institute of Statistics, National Chung Hsing University,* [a] *huynhhuudinh@iuh.edu.vn*
[2] *Department of Mathematics and Statistics, University of Otago,* [b] *matthew.schofield@otago.ac.nz*
[3] *Institute of Statistics, National Tsing Hua University,* [c] *wenhan@stat.nthu.edu.tw*

We propose an enhanced site occupancy model for analyzing ecological detection/nondetection data obtained from multiple visits. The model distinguishes between abundance, occupancy, and detection probabilities. We allow for transient individuals through a community parameter, c, that characterizes the proportion of individuals fixed across visits. This parameter seamlessly transitions from the standard occupancy model ($c = 0$) to the N-mixture model ($c = 1$), enabling a more accurate analysis of site occupancy data. Through theoretical developments and simulation studies, we demonstrate how this model effectively addresses biases inherent in conventional approaches, particularly for c is not at 0 or 1. We apply the model to various datasets of mammal and bird species and compare it to current approaches.

REFERENCES

BARKER, R. J., SCHOFIELD, M. R., LINK, W. A. and SAUER, J. R. (2018). On the reliability of N-mixture models for count data. *Biometrics* **74** 369–377. MR3777957 https://doi.org/10.1111/biom.12734

BROWN, D. D., LAPOINT, S., KAYS, R., HEIDRICH, W., KÜMMETH, F. and WIKELSKI, M. (2012). Accelerometer-informed GPS telemetry: Reducing the trade-off between resolution and longevity. *Wildl. Soc. Bull.* **36** 139–146.

BROWN, J. H. (1984). On the relationship between abundance and distribution of species. *Amer. Nat.* **124** 255–279.

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

DÉNES, F. V., SILVEIRA, L. F. and BEISSINGER, S. R. (2015). Estimating abundance of unmarked animal populations: Accounting for imperfect detection and other sources of zero inflation. *Methods Ecol. Evol.* **6** 543–556.

DENNIS, E. B., MORGAN, B. J. T. and RIDOUT, M. S. (2015). Computational aspects of N-mixture models. *Biometrics* **71** 237–246. MR3335368 https://doi.org/10.1111/biom.12246

DORAZIO, R. M. and ROYLE, J. A. (2005). Estimating size and composition of biological communities by modeling the occurrence of species. *J. Amer. Statist. Assoc.* **100** 389–398. MR2170462 https://doi.org/10.1198/016214505000000015

DROVETSKI, S. V., AGHAYAN, S. A., MATA, V. A., LOPES, R. J., MODE, N. A., HARVEY, J. A. and VOELKER, G. (2014). Does the niche breadth or trade-off hypothesis explain the abundance-occupancy relationship in avian Haemosporidia? *Mol. Ecol.* **23** 3322–3329. https://doi.org/10.1111/mec.12744

DUARTE, A., ADAMS, M. J. and PETERSON, J. T. (2018). Fitting N-mixture models to count data with unmodeled heterogeneity: Bias, diagnostics, and alternative approaches. *Ecol. Model.* **374** 51–59.

GASTON, K. J., BLACKBURN, T. M., GREENWOOD, J. J. D., GREGORY, R. D., QUINN, R. M. and LAWTON, J. H. (2000). Abundance–occupancy relationships. *J. Appl. Ecol.* **37** 39–59.

GOMEZ, J. P., ROBINSON, S. K., BLACKBURN, J. K. and PONCIANO, J. M. (2018). An efficient extension of N-mixture models for multi-species abundance estimation. *Methods Ecol. Evol.* **9** 340–353. https://doi.org/10.1111/2041-210X.12856

HAINES, L. M. (2016a). A note on the Royle–Nichols model for repeated detection-nondetection data. *J. Agric. Biol. Environ. Stat.* **21** 588–598. MR3542088 https://doi.org/10.1007/s13253-016-0253-6

HAINES, L. M. (2016b). Maximum likelihood estimation for N-mixture models. *Biometrics* **72** 1235–1245. MR3591608 https://doi.org/10.1111/biom.12521

Key words and phrases. N-mixture model, occupancy probability, zero-inflated model.

HAYES, D. B. and MONFILS, M. J. (2015). Occupancy modeling of bird point counts: Implications of mobile animals. *J. Wildl. Manag.* **79** 1361–1368.

HE, F. and GASTON, K. J. (2000). Estimating species abundance from occurrence. *Amer. Nat.* **156** 553–559. https://doi.org/10.1086/303403

HE, F. and GASTON, K. J. (2003). Occupancy, spatial variance, and the abundance of species. *Amer. J. Anat.* **162** 366–375. https://doi.org/10.1086/377190

HOGG, S. E., WANG, Y. and STONE, L. (2021). Effectiveness of joint species distribution models in the presence of imperfect detection. *Methods Ecol. Evol.* **12** 1458–1474.

HUI, C., MCGEOCH, M. A., REYERS, B., ROUX, P. C., GREVE, M. and CHOWN, S. L. (2009). Extrapolating population size from the occupancy–abundance relationship and the scaling pattern of occupancy. *Ecol. Appl.* **19** 2038–2048.

HUYNH, H.-D., SCHOFIELD, M. and HWANG, W.-H. (2024). Supplement to "Site occupancy And abundance models For analyzing multiple-visit detection/nondetection data." https://doi.org/10.1214/24-AOAS1888SUPPA, https://doi.org/10.1214/24-AOAS1888SUPPB

HWANG, W. H. and HE, F. (2011). Estimating abundance from presence-absence map. *Methods Ecol. Evol.* **2** 550–559.

JOHNSON, D. S., CONN, P. B., HOOTEN, M. B., RAY, J. C. and POND, B. A. (2013). Spatial occupancy models for large data sets. *Ecology* **94** 801–808.

JORDAN, M. J., BARRETT, R. H. and PURCELL, K. L. (2011). Camera trapping estimates of density and survival of fishers Martes pennanti. *Wildl. Biol.* **17** 266–276.

JOSEPH, L. N., ELKIN, C., MARTIN, T. G. and POSSINGHAM, H. P. (2009). Modeling abundance using N-mixture models: The importance of considering ecological mechanisms. *Ecol. Appl.* **19** 631–642.

KARAVARSAMIS, N. and HUGGINS, R. M. (2020). Two-stage approaches to the analysis of occupancy data I: The homogeneous case (analysis of occupancy data). *Comm. Statist. Theory Methods* **49** 4751–4761. MR4135406 https://doi.org/10.1080/03610926.2019.1607385

KENDALL, W. L., HINES, J. E., NICHOLS, J. D. and GRANT, E. H. C. (2013). Relaxing the closure assumption in occupancy models: Staggered arrival and departure times. *Ecology* **94** 610–617. https://doi.org/10.1890/12-1720.1

KENDALL, W. L. and WHILE, G. C. (2009). A cautionary note on substituting spatial subunits for repeated temporal sampling in studies of site occupancy. *J. Appl. Ecol.* **46** 1182–1188.

KÉRY, M. (2018). Identifiability in N-mixture models: A large-scale screening test with bird data. *Ecology* **99** 281–288. https://doi.org/10.1002/ecy.2093

KNAPE, J., ARLT, D., BARRAQUAND, F., BERG, A., CHEVALIER, M., PÄRT, T., RUETE, A. and ŻMIHORSKI, M. (2018). Sensitivity of binomial N-mixture models to overdispersion: The importance of assessing model fit. *Methods Ecol. Evol.* **9** 2102–2114.

LINK, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* **59** 1123–1130. MR2019822 https://doi.org/10.1111/j.0006-341X.2003.00129.x

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

MACKENZIE, D. I., NICHOLS, J. D., LACHMAN, G. B., DROEGE, S., ROYLE, J. A. and LANGTIMM, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83** 2248–2255.

MACKENZIE, D. I., NICHOLS, J. D., ROYLE, J. A., POLLOCK, K. H., BAILEY, L. L. and HINES, J. E. (2017). *Occupancy Estimation and Modeling*: *Inferring Patterns and Dynamics of Species Occurrence*, 2nd ed. Elsevier, London.

MACKENZIE, D. I., NICHOLS, J. D., SEAMANS, M. E. and GUTIÉRREZ, R. J. (2009). Modeling species occurrence dynamics with multiple states and imperfect detection. *Ecology* **90** 823–835. https://doi.org/10.1890/08-0141.1

MADSEN, L. and ROYLE, J. A. (2023). A review of N-mixture models. *Wiley Interdiscip. Rev.*: *Comput. Stat.* **15** Paper No. e1625, 15. MR4662508

MI, X., BEKERMAN, W., RUSTGI, A. K., SIMS, P. A., CANOLL, P. D. and HU, J. (2024). RZiMM-scRNA: A regularized zero-inflated mixture model framework for single-cell RNA-seq data. *Ann. Appl. Stat.* **18** 1–22. MR4698595 https://doi.org/10.1214/23-aoas1761

NACHMAN, G. (1981). A mathematical model of the functional relationship between density and spatial distribution of a population. *J. Anim. Ecol.* **50** 453–460. MR0646147 https://doi.org/10.2307/4066

OTTO, C. R. V., BAILEY, L. L. and ROLOFF, G. J. (2013). Improving species occupancy estimation when sampling violates the closure assumption. *Ecography* **36** 1299–1309.

PRIYADARSHANI, D., ALTWEGG, R., LEE, A. T. K. and HWANG, W.-H. (2022). What can occupancy models gain from time-to-detection data? *Ecology* **103** e3832. https://doi.org/10.1002/ecy.3832

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

ROSSMAN, S., YACKULIC, C. B., SAUNDERS, S. P., REID, J., DAVIS, R. and ZIPKIN, E. F. (2016). Dynamic N-occupancy models: Estimating demographic rates and local abundance from detection-nondetection data. *Ecology* **97** 3300–3307. https://doi.org/10.1002/ecy.1598

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

ROYLE, J. A. (2006). Site occupancy models with heterogeneous detection probabilities. *Biometrics* **62** 97–102, 316. MR2226561 https://doi.org/10.1111/j.1541-0420.2005.00439.x

ROYLE, J. A. and DORAZIO, R. M. (2008). *Hierarchical Modeling and Inference in Ecology*: *The Analysis of Data from Populations*, *Metapopulations and Communities*. Elsevier, Amsterdam.

ROYLE, J. A. and NICHOLS, J. D. (2003). Estimating abundance from repeated presence-absence data or point counts. *Ecology* **84** 777–790.

SELF, S. G. and LIANG, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82** 605–610. MR0898365

STREBEL, N., FISS, C. J., KELLNER, K. F., LARKIN, J. L., KÉRY, M. and COHEN, J. (2021). Estimating abundance based on time-to-detection data. *Methods Ecol. Evol.* **12** 909–920.

WRIGHT, D. H. (1991). Correlations between incidence and abundance are expected by chance. *J. Biogeogr.* **18** 463–466.

ZIELINSKI, W. J., TRUEX, R. L., SCHLEXER, F. V., CAMPBELL, L. A. and CARROLL, C. (2005). Historical and contemporary distributions of carnivores in forests of the Sierra Nevada. *J. Biogeogr.* **32** 1385–1407.

ZUCKERBERG, B., PORTER, W. F. and CORWIN, K. (2009). The consistency and stability of abundance-occupancy relationships in large-scale population dynamics. *J. Anim. Ecol.* **78** 172–181. https://doi.org/10.1111/j.1365-2656.2008.01463.x

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2444–2461
https://doi.org/10.1214/24-AOAS1889

JOINT MODELING OF MULTISTATE AND NONPARAMETRIC MULTIVARIATE LONGITUDINAL DATA

BY LU YOU[1,a], FALASTIN SALAMI[2,c], CARINA TÖRN[2,d], ÅKE LERNMARK[2,e] AND ROY TAMURA[1,b]

[1]*Health Informatics Institute, University of South Florida,* [a]*lu.you@epi.usf.edu,* [b]*roy.tamura@epi.usf.edu*
[2]*Department of Clinical Sciences, Lund University,* [c]*falastin.salami@med.lu.se,* [d]*carina.torn@med.lu.se,* [e]*ake.lernmark@med.lu.se*

It is oftentimes the case in studies of disease progression that subjects can move into one of several disease states of interest. Multistate models are an indispensable tool to analyze data from such studies. The Environmental Determinants of Diabetes in the Young (TEDDY) is an observational study of at-risk children from birth to onset of type-1 diabetes (T1D) up through the age of 15. A joint model for simultaneous inference of multistate and multivariate nonparametric longitudinal data is proposed to analyze data and answer the research questions brought up in the study. The proposed method allows us to make statistical inferences, test hypotheses, and make predictions about future state occupation in the TEDDY study. The performance of the proposed method is evaluated by simulation studies. The proposed method is applied to the motivating example to demonstrate the capabilities of the method.

REFERENCES

ALAFCHI, B., MAHJUB, H., TAPAK, L., ROSHANAEI, G. and AMIRZARGAR, M. A. (2021). Two-stage joint model for multivariate longitudinal and multistate processes, with application to renal transplantation data. *J. Probab. Stat.* Art. ID 6641602. MR4244007 https://doi.org/10.1155/2021/6641602

ALBERT, P. S. (2019). Shared random parameter models: A legacy of the biostatistics program at the National Heart, Lung and Blood Institute. *Stat. Med.* **38** 501–511. MR3902594 https://doi.org/10.1002/sim.8011

ATKINSON, M. A., EISENBARTH, G. S. and MICHELS, A. W. (2014). Type 1 diabetes. *Lancet* **383** 69–82.

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.

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

BRUMBACK, B. A. and RICE, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Amer. Statist. Assoc.* **93** 961–994. With comments and a rejoinder by the authors. MR1649194 https://doi.org/10.2307/2669837

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

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

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

FERRER, L., RONDEAU, V., DIGNAM, J., PICKLES, T., JACQMIN-GADDA, H. and PROUST-LIMA, C. (2016). Joint modelling of longitudinal and multi-state processes: Application to clinical progressions in prostate cancer. *Stat. Med.* **35** 3933–3948. MR3545618 https://doi.org/10.1002/sim.6972

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

HSIEH, F., TSENG, Y.-K. and WANG, J.-L. (2006). Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics* **62** 1037–1043. MR2297674 https://doi.org/10.1111/j.1541-0420.2006.00570.x

Key words and phrases. Joint modeling, multistate model, spline regression model, type-1 diabetes.

HUANG, H., LI, Y. and GUAN, Y. (2014). Joint modeling and clustering paired generalized longitudinal trajectories with application to cocaine abuse treatment data. *J. Amer. Statist. Assoc.* **109** 1412–1424. MR3293600 https://doi.org/10.1080/01621459.2014.957286

HUANG, X., LI, G. and ELASHOFF, R. M. (2010). A joint model of longitudinal and competing risks survival data with heterogeneous random effects and outlying longitudinal measurements. *Stat. Interface* **3** 185–195. MR2659510 https://doi.org/10.4310/SII.2010.v3.n2.a6

JOY, C., BOYLE, P. P. and TAN, K. S. (1996). Quasi-Monte Carlo methods in numerical finance. *Manage. Sci.* **42** 926–938.

KRISCHER, J. P., LYNCH, K. F., SCHATZ, D. A., ILONEN, J., LERNMARK, Å., HAGOPIAN, W. A., REWERS, M. J., SHE, J.-X., SIMELL, O. G. et al. (2015). The 6 year incidence of diabetes-associated autoantibodies in genetically at-risk children: The TEDDY study. *Diabetologia* **58** 980–987.

LI, G., LESPERANCE, M. and WU, Z. (2022). Joint modeling of multivariate survival data with an application to retirement. *Sociol. Methods Res.* **51** 1920–1946. MR4516875 https://doi.org/10.1177/0049124120914928

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

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

MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. With comments and a rejoinder by the authors. MR1803168 https://doi.org/10.2307/2669386

NIEDERREITER, H. (1978). Quasi-Monte Carlo methods and pseudo-random numbers. *Bull. Amer. Math. Soc.* **84** 957–1041. MR0508447 https://doi.org/10.1090/S0002-9904-1978-14532-7

PAK, D., LI, C., TODEM, D. and SOHN, W. (2017). A multistate model for correlated interval-censored life history data in caries research. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 413–423. MR3611694 https://doi.org/10.1111/rssc.12186

PERPEROGLOU, A., SAUERBREI, W., ABRAHAMOWICZ, M. and SCHMID, M. (2019). A review of spline function procedures in R. *BMC Med. Res. Methodol.* **19** 1–16.

RIZOPOULOS, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67** 819–829. MR2829256 https://doi.org/10.1111/j.1541-0420.2010.01546.x

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

SALAMI, F., TAMURA, R., YOU, L., LERNMARK, Å., LARSSON, H. E., LUNDGREN, M., KRISCHER, J., ZIEGLER, A.-G., TOPPARI, J. et al. (2022). HbA1c as a time predictive biomarker for an additional islet autoantibody and type 1 diabetes in seroconverted TEDDY children. *Pediatric Diabetes* **23** 1586–1593.

SAYERS, A., HERON, J., SMITH, A. D. A. C., MACDONALD-WALLIS, C., GILTHORPE, M. S., STEELE, F. and TILLING, K. (2017). Joint modelling compared with two stage methods for analysing longitudinal data and prospective outcomes: A simulation study of childhood growth and BP. *Stat. Methods Med. Res.* **26** 437–452. MR3592734 https://doi.org/10.1177/0962280214548822

SCHLUCHTER, M. D. (1992). Methods for the analysis of informatively censored longitudinal data. *Stat. Med.* **11** 1861–1870. https://doi.org/10.1002/sim.4780111408

SLOAN, I. H. and WOŹNIAKOWSKI, H. (1998). When are quasi-Monte Carlo algorithms efficient for high-dimensional integrals? *J. Complexity* **14** 1–33. MR1617765 https://doi.org/10.1006/jcom.1997.0463

SOBOL, I. M. (1976). Uniformly distributed sequences with an additional property of uniformity. *USSR Comput. Math. Math. Phys.* **16** 236–242.

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

WAND, M. P. (2000). A comparison of regression spline smoothing procedures. *Comput. Statist.* **15** 443–462. MR1818029 https://doi.org/10.1007/s001800000047

WILLIAMSON, P. R., KOLAMUNNAGE-DONA, R., PHILIPSON, P. and MARSON, A. G. (2008). Joint modelling of longitudinal and competing risks data. *Stat. Med.* **27** 6426–6438. MR2655125 https://doi.org/10.1002/sim.3451

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

XU, J. and ZEGER, S. L. (2001). The evaluation of multiple surrogate endpoints. *Biometrics* **57** 81–87. MR1833292 https://doi.org/10.1111/j.0006-341X.2001.00081.x

YANG, L., YU, M. and GAO, S. (2016). Joint models for multiple longitudinal processes and time-to-event outcome. *J. Stat. Comput. Simul.* **86** 3682–3700. MR3547952 https://doi.org/10.1080/00949655.2016.1181760

YAO, F. (2007). Functional principal component analysis for longitudinal and survival data. *Statist. Sinica* **17** 965–983. MR2408647

Ye, J., Li, Y. and Guan, Y. (2015). Joint modeling of longitudinal drug using pattern and time to first relapse in cocaine dependence treatment data. *Ann. Appl. Stat.* **9** 1621–1642. MR3418738 https://doi.org/10.1214/15-AOAS852

Yiu, S. and Tom, B. (2017). A joint modelling approach for multistate processes subject to resolution and under intermittent observations. *Stat. Med.* **36** 496–508. MR3592175 https://doi.org/10.1002/sim.7149

You, L. and Qiu, P. (2021). Joint modeling of multivariate nonparametric longitudinal data and survival data: A local smoothing approach. *Stat. Med.* **40** 6689–6706. MR4352762 https://doi.org/10.1002/sim.9206

You, L., Salami, F., Törn, C., Lernmark, Å. and Tamura, R. (2024). Supplement to "Joint modeling of multistate and nonparametric multivariate longitudinal data." https://doi.org/10.1214/24-AOAS1889SUPPA, https://doi.org/10.1214/24-AOAS1889SUPPB

Yu, L., Zhao, Z. and Steck, A. K. (2017). T1D autoantibodies: Room for improvement? *Current Opinion in Endocrinology, Diabetes, and Obesity* **24** 285.

Yue, X. and Al Kontar, R. (2021). Joint models for event prediction from time series and survival data. *Technometrics* **63** 477–486. MR4331448 https://doi.org/10.1080/00401706.2020.1832582

Zeng, D. and Lin, D. Y. (2021). Maximum likelihood estimation for semiparametric regression models with panel count data. *Biometrika* **108** 947–963. MR4341361 https://doi.org/10.1093/biomet/asaa091

Zhang, H., Kelvin, E. A., Carpio, A. and Allen Hauser, W. (2020). A multistate joint model for interval-censored event-history data subject to within-unit clustering and informative missingness, with application to neurocysticercosis research. *Stat. Med.* **39** 3195–3206. MR4151928 https://doi.org/10.1002/sim.8663

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2462–2485
https://doi.org/10.1214/24-AOAS1890
© Institute of Mathematical Statistics, 2024

LATENT LEVEL CORRELATION MODELING OF MULTIVARIATE DISCRETE-VALUED FINANCIAL TIME SERIES

BY YANZHAO WANG[1,a], HAITAO LIU[2,c], JIAN ZOU[1,b] AND NALINI RAVISHANKER[3,d]

[1] *Department of Mathmatical Sciences, Worcester Polytechnic Institute,* [a]*ywang34@wpi.edu,* [b]*jzou@wpi.edu*
[2] *Data Science Program, Worcester Polytechnic Institute,* [c]*hliu5@wpi.edu*
[3] *Department of Statistics, University or Connecticut,* [d]*nalini.ravishanker@uconn.edu*

In high-frequency financial data, dynamic patterns of transaction counts in regular time intervals provide crucial insights into market microstructure, such as short-term trading activities and intermittent intensities of price oscillation. In this paper we propose a Bayesian hierarchical framework that incorporates correlated latent level and temporal effects to model multivariate count data during intraday transaction intervals. Built on the INLA method for implementation, our framework proves to be competitive with the traditional MCMC approach in terms of model inference and computational cost. We demonstrate the efficacy of our methodology by applying it to assets from three Global Industry Classification Standard (GICS) sectors, namely, healthcare, energy, and industrials. The analysis uncovers various microstructures of financial count data using our framework. Specifically, our model featuring a correlated latent effect structure adeptly captures the pattern of the empirical correlations within the count data patterns with additional statistical inference, such as assessing different associations between short-term averaged trading size as well as trading duration, the counts at different risk levels, and uncovering differential levels of uncertainty resulted from market temporal behavior and unobservable latent effects across the three sectors. We also discuss some potential applications of our framework in real-world scenarios.

REFERENCES

AITCHISON, J. and HO, C.-H. (1989). The multivariate Poisson-log normal distribution. *Biometrika* **76** 643–653. MR1041409 https://doi.org/10.1093/biomet/76.4.643

AKTEKIN, T., POLSON, N. and SOYER, R. (2018). Sequential Bayesian analysis of multivariate count data. *Bayesian Anal.* **13** 385–409. MR3780428 https://doi.org/10.1214/17-BA1054

AL-OSH, M. A. and ALZAID, A. A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *J. Time Series Anal.* **8** 261–275. MR0903755 https://doi.org/10.1111/j.1467-9892.1987.tb00438.x

BARON, M., BROGAARD, J., HAGSTRÖMER, B. and KIRILENKO, A. (2019). Risk and return in high-frequency trading. *J. Financ. Quant. Anal.* **54** 993–1024.

BERAHA, M., FALCO, D. and GUGLIELMI, A. (2021). JAGS, NIMBLE, Stan: A detailed comparison among Bayesian MCMC software. arXiv preprint. Available at arXiv:2107.09357.

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

BROGAARD, J., CARRION, A., MOYAERT, T., RIORDAN, R., SHKILKO, A. and SOKOLOV, K. (2018). High frequency trading and extreme price movements. *J. Financ. Econ.* **128** 253–265.

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

DE CARVALHO, M., HUSER, R. and RUBIO, R. (2023). Similarity-based clustering for patterns of extreme values. *Stat* **12** e560. MR4604508 https://doi.org/10.1002/sta4.560

EFRON, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* **81** 709–721. MR0860505

Key words and phrases. Bayesian hierarchical model, intraday multivariate financial count data, integrated nested Laplace approximation (INLA), level correlation model (LCM), correlated temporal effects.

FERLAND, R., LATOUR, A. and ORAICHI, D. (2006). Integer-valued GARCH process. *J. Time Series Anal.* **27** 923–942. MR2328548 https://doi.org/10.1111/j.1467-9892.2006.00496.x

GAMERMAN, D., DOS SANTOS, T. R. and FRANCO, G. C. (2013). A non-Gaussian family of state-space models with exact marginal likelihood. *J. Time Series Anal.* **34** 625–645. MR3127211 https://doi.org/10.1111/jtsa.12039

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

HEINEN, A. (2003). Modelling time series count data: An autoregressive conditional Poisson model. Available at SSRN 1117187.

JACOBS, P. A. and LEWIS, P. A. W. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *J. Time Series Anal.* **4** 19–36. MR0711293 https://doi.org/10.1111/j.1467-9892.1983.tb00354.x

JUNG, R. C., LIESENFELD, R. and RICHARD, J.-F. (2011). Dynamic factor models for multivariate count data: An application to stock-market trading activity. *J. Bus. Econom. Statist.* **29** 73–85. MR2789392 https://doi.org/10.1198/jbes.2009.08212

KARLIS, D. and MELIGKOTSIDOU, L. (2005). Multivariate Poisson regression with covariance structure. *Stat. Comput.* **15** 255–265. MR2205389 https://doi.org/10.1007/s11222-005-4069-4

KARLIS, D. and MELIGKOTSIDOU, L. (2007). Finite mixtures of multivariate Poisson distributions with application. *J. Statist. Plann. Inference* **137** 1942–1960. MR2323875 https://doi.org/10.1016/j.jspi.2006.07.001

LAVINE, I., CRON, A. and WEST, M. (2022). Bayesian computation in dynamic latent factor models. *J. Comput. Graph. Statist.* **31** 651–665. MR4495701 https://doi.org/10.1080/10618600.2021.2021208

LIESENFELD, R., NOLTE, I. and POHLMEIER, W. (2006). Modelling financial transaction price movements: A dynamic integer count data model. *Empir. Econ.* **30** 795–825.

LONGIN, F. (2016). *Extreme Events in Finance: A Handbook of Extreme Value Theory and Its Applications. Wiley Handbook in Financial Engineering and Econometrics*. Wiley, New York.

MA, J., KOCKELMAN, K. M. and DAMIEN, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Anal. Prev.* **40** 964–975.

PARK, E. S. and LORD, D. (2007). Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. *Transp. Res. Rec.* **2019** 1–6.

PEDELI, X. and KARLIS, D. (2013). On estimation of the bivariate Poisson INAR process. *Comm. Statist. Simulation Comput.* **42** 514–533. MR3020084 https://doi.org/10.1080/03610918.2011.639001

QUORESHI, A. M. M. S. (2017). A bivariate integer-valued long-memory model for high-frequency financial count data. *Comm. Statist. Theory Methods* **46** 1080–1089. MR3565611 https://doi.org/10.1080/03610926.2014.997361

RAMAN, B., RAVISHANKER, N., SOYER, R., GORTI, V. and SEN, K. (2020). Dynamic Bayesian modeling of multiple count time series using `R-INLA`. *J. Indian Statist. Assoc.* **58** 137–173. MR4361464

RAVISHANKER, N., RAMAN, B. and SOYER, R. (2022). *Dynamic Time Series Models Using R-INLA: An Applied Perspective*. CRC Press, Boca Raton.

RAVISHANKER, N., SERHIYENKO, V. and WILLIG, M. R. (2014). Hierarchical dynamic models for multivariate times series of counts. *Stat. Interface* **7** 559–570. MR3302382 https://doi.org/10.4310/SII.2014.v7.n4.a11

RIEBLER, A. and HELD, L. (2017). Projecting the future burden of cancer: Bayesian age-period-cohort analysis with integrated nested Laplace approximations. *Biom. J.* **59** 531–549. MR3648612 https://doi.org/10.1002/bimj.201500263

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields: Theory and Applications. Monographs on Statistics and Applied Probability* **104**. CRC Press, 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. and LINDGREN, F. K. (2017). Bayesian computing with INLA: A review. *Annu. Rev. Stat. Appl.* **4** 395–421. MR3634300 https://doi.org/10.1214/16-STS576

RUIZ-CÁRDENAS, R., KRAINSKI, E. T. and RUE, H. (2012). Direct fitting of dynamic models using integrated nested Laplace approximations—INLA. *Comput. Statist. Data Anal.* **56** 1808–1828. MR2892379 https://doi.org/10.1016/j.csda.2011.10.024

SADYKOVA, D., SCOTT, B. E., DOMINICIS, M. D., WAKELIN, S. L., SADYKOV, A. and WOLF, J. (2017). Bayesian joint models with INLA exploring marine mobile predator-prey and competitor species habitat overlap. *Ecol. Evol.* **7** 5212–5226. https://doi.org/10.1002/ece3.3081

SALMON, M., SCHUMACHER, D., STARK, K. and HÖHLE, M. (2015). Bayesian outbreak detection in the presence of reporting delays. *Biom. J.* **57** 1051–1067. MR3415359 https://doi.org/10.1002/bimj.201400159

SCHRÖDLE, B. and HELD, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics* **22** 725–734. MR2843139 https://doi.org/10.1002/env.1065

SERHIYENKO, V., RAVISHANKER, N. and VENKATESAN, R. (2018). Multi-stage multivariate modeling of temporal patterns in prescription counts for competing drugs in a therapeutic category. *Appl. Stoch. Models Bus. Ind.* **34** 61–78. MR3769490 https://doi.org/10.1002/asmb.2232

SOYER, R. and ZHANG, D. (2022). Bayesian modeling of multivariate time series of counts. *Wiley Interdiscip. Rev.: Comput. Stat.* **14** Paper No. e1559, 18. MR4515042

WANG, Y., LIU, H., ZOU, J. and RAVISHANKER, N. (2024). Supplement to "Latent level correlation modeling of multivariate discrete-valued financial time series." https://doi.org/10.1214/24-AOAS1890SUPP

WATANABE, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11** 3571–3594. MR2756194

WEST, M. (2020). Bayesian forecasting of multivariate time series: Scalability, structure uncertainty and decisions. *Ann. Inst. Statist. Math.* **72** 1–31. MR4052647 https://doi.org/10.1007/s10463-019-00741-3

WEST, M., HARRISON, P. J. and MIGON, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *J. Amer. Statist. Assoc.* **80** 73–83. MR0786598

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2486–2510
https://doi.org/10.1214/24-AOAS1891
© Institute of Mathematical Statistics, 2024

A FORENSIC STATISTICAL ANALYSIS OF FRAUD IN THE FEDERAL FOOD STAMP PROGRAM

BY JONATHAN WOODY[1,a], ZHICONG ZHAO[1,b], ROBERT LUND[2,d] AND TUNG-LUNG WU[1,c]

[1]*Department of Mathematics and Statistics, Mississippi State University,* [a]*jwoody@math.msstate.edu,* [b]*zz204@msstate.edu,* [c]*twu@math.msstate.edu*

[2]*Department of Statistics, University of California, Santa Cruz,* [d]*rolund@ucsc.edu*

This study develops methods to detect anomalous transactions linked with fraud in food stamp purchases through order statistics methods. The methods detect clusters in the order statistics of the transaction amounts that merit further scrutiny. Our techniques use scan statistics to determine when an excessive number of transactions occur (cluster), which is historically linked to fraud. A scoring paradigm is constructed that ranks the degree in which detected clusters and individual transactions are anomalous among approximately 250 million total transactions.

REFERENCES

ABDALLAH, A., MAAROF, M. A. and ZAINAL, A. (2016). Fraud detection system: A survey. *J. Netw. Comput. Appl.* **68** 90–113.

AGGARWAL, C. C. (2015). Outlier analysis. In *Data Mining* 237–263. Springer.

AGGARWAL, C. C., ed. (2015). *Data Classification: Algorithms and Applications*. CRC press, Boca Raton.

AHMED, M., MAHMOOD, A. N. and ISLAM, M. R. (2016). A survey of anomaly detection techniques in financial domain. *Future Gener. Comput. Syst.* **55** 278–288.

AHSANULLAH, M., NEVZOROV, V. B. and SHAKIL, M. (2013). *An Introduction to Order Statistics*. *Atlantis Studies in Probability and Statistics* **3**. Atlantis Press, Paris. MR3025012 https://doi.org/10.2991/978-94-91216-83-1

AL-HASHEDI, K. G. and MAGALINGAM, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Comput. Sci. Rev.* **40** 100402.

ARNOLD, B. C., BALAKRISHNAN, N. and NAGARAJA, H. N. (2008). *A First Course in Order Statistics*. *Classics in Applied Mathematics* **54**. SIAM, Philadelphia, PA. MR2399836 https://doi.org/10.1137/1.9780898719062

BOLTON, R. J. and HAND, D. J. (2002). Statistical fraud detection: A review. *Statist. Sci.* **17** 235–249. MR1963313 https://doi.org/10.1214/ss/1042727940

CANNING, P. and STACY, B. (2019). The Supplemental Nutrition Assistance Program (SNAP) and the economy: New estimates of the SNAP multiplier. Technical Report.

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

CLINE, D. R. and AUSSENBERG, R. A. (2018). *Errors and Fraud in the Supplemental Nutrition Assistance Program (SNAP)*. (CRS Report No. R45147).

COUNCIL, N. R. et al. (2013). *Supplemental Nutrition Assistance Program: Examining the Evidence to Define Benefit Adequacy*. National Academies Press.

DAVID, H. A. and NAGARAJA, H. N. (2003). *Order Statistics*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1994955 https://doi.org/10.1002/0471722162

DEAN, S. (2016). *SNAP: Combating Fraud and Improving Program Integrity Without Weakening Success: Hearings Before the Subcommittees on Government Operations and the Interior of the Committee on Oversight and Goverment Reform U.S. House of Represntatives*. 114th Congress.

DURTSCHI, C., HILLISON, W. and PACINI, C. (2004). The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting* **5** 17–34.

EKIN, T., IEVA, F., RUGGERI, F. and SOYER, R. (2018). Statistical medical fraud assessment: Exposition to an emerging field. *Int. Stat. Rev.* **86** 379–402. MR3882123 https://doi.org/10.1111/insr.12269

FAULK, K. (2016). Alabama grocier to forfiet $5.2 million in food stamp fraud case. *Birmiham Real-Time News*.

Key words and phrases. Order statistics, fraud detection, scan statistics, data science, SNAP trafficking.

FEIGIN, P. D. (1979). On the characterization of point processes with the order statistic property. *J. Appl. Probab.* **16** 297–304. MR0531764 https://doi.org/10.1017/s0021900200046507

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications. Vol. II*, 1st ed. Wiley, New York. MR0210154

FU, J. C. (2001). Distribution of the scan statistic for a sequence of bistate trials. *J. Appl. Probab.* **38** 908–916. MR1876548 https://doi.org/10.1017/s0021900200019124

GOOD, I. J. and GASKINS, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75** 42–73. MR0568579

HAIMAN, G. (2007). Estimating the distribution of one-dimensional discrete scan statistics viewed as extremes of 1-dependent stationary sequences. *J. Statist. Plann. Inference* **137** 821–828. MR2301718 https://doi.org/10.1016/j.jspi.2006.06.010

HAND, D. J., BLUNT, G., KELLY, M. G., ADAMS, N. M. et al. (2000). Data mining for fun and profit. *Statist. Sci.* **15** 111–131.

HE, Z., XU, X., HUANG, Z. J. and DENG, S. (2005). FP-outlier: Frequent pattern based outlier detection. *Comput. Sci. Inf. Syst.* **2** 103–118.

HILAS, C. S. and MASTOROCOSTAS, P. A. (2008). An application of supervised and unsupervised learning approaches to telecommunications fraud detection. *Knowl.-Based Syst.* **21** 721–726.

KARLIN, S. and TAYLOR, H. M. (1981). *A Second Course in Stochastic Processes*. Academic Press, New York. MR0611513

KONIJN, R. M., DUIVESTEIJN, W., KOWALCZYK, W. and KNOBBE, A. (2013). Discovering local subgroups, with an application to fraud detection. In *Advances in Knowledge Discovery and Data Mining* (J. Pei, V. S. Tseng, L. Cao, H. Motoda and G. Xu, eds.) 1–12. Springer, Berlin.

KOU, Y., LU, C.-T., SIRWONGWATTANA, S. and HUANG, Y.-P. (2004). Survey of fraud detection techniques. In *IEEE International Conference on Networking, Sensing and Control*, 2004 **2** 749–754. IEEE, New York.

LIAO, H.-J., LIN, C.-H. R., LIN, Y.-C. and TUNG, K.-Y. (2013). Intrusion detection system: A comprehensive review. *J. Netw. Comput. Appl.* **36** 16–24.

LIBERMAN, U. (1985). An order statistic characterization of the Poisson renewal process. *J. Appl. Probab.* **22** 717–722. MR0799295 https://doi.org/10.1017/s0021900200029478

LIU, X. and ZHANG, P. (2010). A scan statistics based suspicious transactions detection model for anti-money laundering (AML) in financial institutions. In 2010 *International Conference on Multimedia Communications* 210–213.

MAES, S., TUYLS, K., VANSCHOENWINKEL, B. and MANDERICK, B. (2002). Credit card fraud detection using Bayesian and neural networks. In *Proceedings of the* 1*st International Naiso Congress on Neuro Fuzzy Technologies* 261–270.

MILLER, S. J., ed. (2015). *Benford's Law*. Princeton Univ. Press, Princeton, NJ. MR3408774 https://doi.org/10.1515/9781400866595

NAGLER, T. (2018). Asymptotic analysis of the jittering kernel density estimator. *Math. Methods Statist.* **27** 32–46. MR3800980 https://doi.org/10.3103/S1066530718010027

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

NGAI, E. W., HU, Y., WONG, Y. H., CHEN, Y. and SUN, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decis. Support Syst.* **50** 559–569.

PHUA, C., LEE, V., SMITH-MILES, K. and GAYLER, R. (2010). A comprehensive survey of data mining-based fraud detection research. *CoRR*. Available at arXiv:1009.6119.

RAJ, S. B. E. and PORTIA, A. A. (2011). Analysis on credit card fraud detection methods. In 2011 *International Conference on Computer, Communication and Electrical Technology* (*ICCCET*) 152–156. IEEE, New York.

SCOTT, D. W. (1979). On optimal and data-based histograms. *Biometrika* **66** 605–610. MR0556742 https://doi.org/10.1093/biomet/66.3.605

SHAO, M., LI, J., CHANG, Y., ZHAO, J. and CHEN, X. (2021). MASA: An efficient framework for anomaly detection in multi-attributed networks. *Computers & Security* **102** 102085.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability*. CRC Press, London. MR0848134 https://doi.org/10.1007/978-1-4899-3324-9

STIRZAKER, D. (2003). *Elementary Probability*, 2nd ed. Cambridge Univ. Press, Cambridge. MR1998578 https://doi.org/10.1017/CBO9780511755309

THIPRUNGSRI, S. and VASARHELYI, M. A. (2011). Cluster analysis for anomaly detection in accounting data: An audit approach. *Int. J. Digit. Account. Res.* **11** 69–84.

TORGO, L. (2011). *Data Mining with R: Learning with Case Studies*. CRC Press/CRC, Boca Raton.

WAND, M. (1997). Data-based choice of histogram bin width. *Amer. Statist.* **51** 59–64.

WATSON, G. S. and LEADBETTER, M. R. (1964). Hazard analysis. I. *Biometrika* **51** 175–184. MR0184335 https://doi.org/10.1093/biomet/51.1-2.175

WILSON, H. (2017). The extent of trafficking in the supplemental nutrition assistance program: 2012-2014. *Nutrition Assistance Program Report*.

WOODY, J., ZHAO, Z., LUND, R. and WU, T.-L. (2024). Supplement to "A forensic statistical analysis of fraud in the federal food stamp program." https://doi.org/10.1214/24-AOAS1891SUPP

WU, Q. and GLAZ, J. (2019). Robust scan statistics for detecting a local change in population mean for normal data. *Methodol*. *Comput*. *Appl*. *Probab*. **21** 295–314. MR3915443 https://doi.org/10.1007/s11009-018-9668-6

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2511–2531
https://doi.org/10.1214/24-AOAS1892

BAYESIAN SPARSE VECTOR AUTOREGRESSIVE SWITCHING MODELS WITH APPLICATION TO HUMAN GESTURE PHASE SEGMENTATION

BY BENIAMINO HADJ-AMAR[1,a], JACK JEWSON[2,c] AND MARINA VANNUCCI[1,b]

[1] *Department of Statistics, Rice University,* [a]*bh44@rice.edu,* [b]*marina@rice.edu*
[2] *Department of Economics and Business, Universitat Pompeu Fabra,* [c]*jack.jewson@upf.edu*

We propose a sparse vector autoregressive (VAR) hidden semi-Markov model (HSMM) for modeling temporal and contemporaneous (e.g., spatial) dependencies in multivariate nonstationary time series. The HSMM's generic state distribution is embedded in a special transition matrix structure, facilitating efficient likelihood evaluations and arbitrary approximation accuracy. To promote sparsity of the VAR coefficients, we deploy an l_1-ball projection prior, which combines differentiability with a positive probability of obtaining exact zeros, achieving variable selection within each switching state. This also facilitate posterior estimation via HMC. We further place nonlocal priors on the parameters of the HSMM dwell distribution improving the ability of Bayesian model selection to distinguish whether the data is better supported by the simpler hidden Markov model (HMM) or more flexible HSMM. Our proposed methodology is illustrated via an application to human gesture phase segmentation based on sensor data, where we successfully identify and characterize the periods of rest and active gesturing as well as the dynamical patterns involved in the gesture movements associated with each of these states.

REFERENCES

AHELEGBEY, D. F., BILLIO, M. and CASARIN, R. (2016). Bayesian graphical models for structural vector autoregressive processes. *J. Appl. Econometrics* **31** 357–386. MR3481367 https://doi.org/10.1002/jae.2443

ALLEN, E. A., DAMARAJU, E., PLIS, S. M., ERHARDT, E. B., EICHELE, T. and CALHOUN, V. D. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cereb. Cortex* **24** 663–676.

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

BENSON, A. and FRIEL, N. (2021). Bayesian inference, model selection and likelihood estimation using fast rejection sampling: The Conway-Maxwell-Poisson distribution. *Bayesian Anal.* **16** 905–931. MR4303873 https://doi.org/10.1214/20-BA1230

BILLIO, M., CASARIN, R. and ROSSINI, L. (2019). Bayesian nonparametric sparse VAR models. *J. Econometrics* **212** 97–115. MR3994009 https://doi.org/10.1016/j.jeconom.2019.04.022

BRIER, G. W. et al. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78** 1–3.

CARPENTER, B., GELMAN, A., HOFFMAN, M., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M. A., GUO, J., LI, P. et al. (2016). Stan: A probabilistic programming language. *J. Stat. Softw.* **20**.

CHIANG, S., GUINDANI, M., YEH, H. J., HANEEF, Z., STERN, J. M. and VANNUCCI, M. (2017). Bayesian vector autoregressive model for multi-subject effective connectivity inference using multi-modal neuroimaging data. *Hum. Brain Mapp.* **38** 1311–1332. https://doi.org/10.1002/hbm.23456

CHICCO, D. and JURMAN, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21** 1–13.

CONWAY, R. W. and MAXWELL, W. L. (1962). A queuing model with state dependent service rates. *J. Ind. Eng.* **12** 132–136.

DOAN, T., LITTERMAN, R. and SIMS, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Rev.* **3** 1–100.

DUANE, S., KENNEDY, A. D., PENDLETON, B. J. and ROWETH, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195** 216–222. MR3960671 https://doi.org/10.1016/0370-2693(87)91197-x

Key words and phrases. Hidden semi-Markov models, vector autoregressive, sparsity, switching models, gesture phase segmentation.

DUCHI, J., SHALEV-SHWARTZ, S., SINGER, Y. and CHANDRA, T. (2008). Efficient projections onto the l_1-ball for learning in high dimensions. In *Proceedings of the* 25*th International Conference on Machine Learning* 272–279.

FOX, E. B., HUGHES, M. C., SUDDERTH, E. B. and JORDAN, M. I. (2014). Joint modeling of multiple time series via the beta process with application to motion capture segmentation. *Ann. Appl. Stat.* **8** 1281–1313. MR3271333 https://doi.org/10.1214/14-AOAS742

FÚQUENE, J., STEEL, M. and ROSSELL, D. (2019). On choosing mixture components via non-local priors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 809–837. MR4025398 https://doi.org/10.1111/rssb.12333

GEFANG, D. (2014). Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage. *Int. J. Forecast.* **30** 1–11.

GELMAN, A. and HILL, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GHOSH, S., KHARE, K. and MICHAILIDIS, G. (2021). Strong selection consistency of Bayesian vector autoregressive models based on a pseudo-likelihood approach. *Ann. Statist.* **49** 1267–1299. MR4298864 https://doi.org/10.1214/20-aos1992

GOEBEL, R., ROEBROECK, A., KIM, D.-S. and FORMISANO, E. (2003). Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. *Magn. Reson. Imaging* **21** 1251–1261.

GRONAU, Q., SINGMANN, H. and WAGENMAKERS, E.-J. (2020). Bridgesampling: An R package for estimating normalizing constants. *J. Stat. Softw.* **92**.

GUÉDON, Y. (2003). Estimating hidden semi-Markov chains from discrete sequences. *J. Comput. Graph. Statist.* **12** 604–639. MR2002638 https://doi.org/10.1198/1061860032030

HADJ-AMAR, B., FINKENSTÄDT, B., FIECAS, M. and HUCKSTEPP, R. (2021). Identifying the recurrence of sleep apnea using a harmonic hidden Markov model. *Ann. Appl. Stat.* **15** 1171–1193. MR4316645 https://doi.org/10.1214/21-aoas1455

HADJ-AMAR, B., JEWSON, J. and FIECAS, M. (2023). Bayesian approximations to hidden semi-Markov models for telemetric monitoring of physical activity. *Bayesian Anal.* **18** 547–577. MR4578064 https://doi.org/10.1214/22-ba1318

HADJ-AMAR, B., JEWSON, J. and VANNUCCI, M. (2024). Supplement to "Bayesian sparse vector autoregressive switching models with application to human gesture phase segmentation." https://doi.org/10.1214/24-AOAS1892SUPPA, https://doi.org/10.1214/24-AOAS1892SUPPB

HAMILTON, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57** 357–384. MR0996941 https://doi.org/10.2307/1912559

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

HUANG, Q., COHEN, D., KOMARZYNSKI, S., LI, X.-M., INNOMINATO, P., LÉVI, F. and FINKENSTÄDT, B. (2018). Hidden Markov models for monitoring circadian rhythmicity in telemetric activity data. *J. R. Soc. Interface* **15** 20170885.

JOHNSON, V. E. and ROSSELL, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. MR2980074 https://doi.org/10.1080/01621459.2012.682536

KADIYALA, K. R. and KARLSSON, S. (1997). Numerical methods for estimation and inference in Bayesian VAR-models. *J. Appl. Econometrics* **12** 99–132.

KALLI, M. and GRIFFIN, J. E. (2018). Bayesian nonparametric vector autoregressive models. *J. Econometrics* **203** 267–282. MR3770826 https://doi.org/10.1016/j.jeconom.2017.11.009

KAMMERDINER, A. R. and PARDALOS, P. M. (2010). Analysis of multichannel EEG recordings based on generalized phase synchronization and cointegrated VAR. In *Computational Neuroscience*. *Springer Optim. Appl.* **38** 317–339. Springer, New York. MR2656819 https://doi.org/10.1007/978-0-387-88630-5_18

LANGROCK, R. and ZUCCHINI, W. (2011). Hidden Markov models with arbitrary state dwell-time distributions. *Comput. Statist. Data Anal.* **55** 715–724. MR2736590 https://doi.org/10.1016/j.csda.2010.06.015

LEROUX, B. G. and PUTERMAN, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* **48** 545–558.

LEWANDOWSKI, D., KUROWICKA, D. and JOE, H. (2009). Generating random correlation matrices based on Vines and extended onion method. *J. Multivariate Anal.* **100** 1989–2001. MR2543081 https://doi.org/10.1016/j.jmva.2009.04.008

LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin. MR2172368 https://doi.org/10.1007/978-3-540-27752-1

MADEO, R. C., LIMA, C. A. and PERES, S. M. (2013). Gesture unit segmentation using support vector machines: Segmenting gestures from rest positions. In *Proceedings of the* 28*th Annual ACM Symposium on Applied Computing* 46–52.

MENG, X.-L. and SCHILLING, S. (2002). Warp bridge sampling. *J. Comput. Graph. Statist.* **11** 552–586. MR1938446 https://doi.org/10.1198/106186002457

MENG, X.-L. and WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860. MR1422406

MITRA, S. and ACHARYA, T. (2007). Gesture recognition: A survey. *IEEE Trans. Syst. Man Cybern., Part C Appl. Rev.* **37** 311–324.

MONI, M. and ALI, A. S. (2009). HMM based hand gesture recognition: A review on techniques and approaches. In 2009 *2nd IEEE International Conference on Computer Science and Information Technology* 433–437. IEEE.

OMBAO, H., FIECAS, M., TING, C.-M. and LOW, Y. F. (2018). Statistical models for brain signals with properties that evolve across trials. *NeuroImage* **180** 609–618. https://doi.org/10.1016/j.neuroimage.2017.11.061

PACI, L. and CONSONNI, G. (2020). Structural learning of contemporaneous dependencies in graphical VAR models. *Comput. Statist. Data Anal.* **144** 106880. MR4030320 https://doi.org/10.1016/j.csda.2019.106880

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

PARVATHY, P., SUBRAMANIAM, K., PRASANNA VENKATESAN, G., KARTHIKAIKUMAR, P., VARGHESE, J. and JAYASANKAR, T. (2021). Development of hand gesture recognition system using machine learning. *J. Ambient Intell. Humaniz. Comput.* **12** 6793–6800.

PFENNINGER, S. (2017). Dealing with multiple decades of hourly wind and PV time series in energy models: A comparison of methods to reduce time resolution and the planning implications of inter-annual variability. *Appl. Energy* **197** 1–13.

PIMENTEL, M. A., SANTOS, M. D., SPRINGER, D. B. and CLIFFORD, G. D. (2015). Heart beat detection in multimodal physiological data using a hidden semi-Markov model and signal quality indices. *Physiol. Meas.* **36** 1717.

PRADO, R., MOLINA, F. and HUERTA, G. (2006). Multivariate time series modeling and classification via hierarchical VAR mixtures. *Comput. Statist. Data Anal.* **51** 1445–1462. MR2307518 https://doi.org/10.1016/j.csda.2006.03.002

ROMANUKE, V. (2021). Time series smoothing improving forecasting. *Appl. Comput. Syst.* **26** 60–70.

ROSSELL, D. and TELESCA, D. (2017). Nonlocal priors for high-dimensional estimation. *J. Amer. Statist. Assoc.* **112** 254–265. MR3646569 https://doi.org/10.1080/01621459.2015.1130634

SAMDIN, S. B., TING, C.-M., OMBAO, H. and SALLEH, S.-H. (2017). A unified estimation framework for state-related changes in effective brain connectivity. *IEEE Trans. Biomed. Eng.* **64** 844–858. https://doi.org/10.1109/TBME.2016.2580738

SARKAR, A., HOSSAIN, S. S. and SARKAR, R. (2023). Human activity recognition from sensor data using spatial attention-aided CNN with genetic algorithm. *Neural Comput. Appl.* **35** 5165–5191.

SIMS, C. A. (1980). Macroeconomics and reality. *Econometrica* 1–48.

TEAM, S. D. (2018). Stan User's Guide.

VINCIOTTI, V., BEHROUZI, P. and MOHAMMADI, R. (2022). Bayesian structural learning of microbiota systems from count metagenomic data. ArXiv preprint. Available at arXiv:2203.10118.

WAGNER, P. K., PERES, S. M., MADEO, R. C. B., DE MORAES LIMA, C. A. and DE ALMEIDA FREITAS, F. (2014). Gesture unit segmentation using spatial-temporal information and machine learning. In *The Twenty-Seventh International Flairs Conference*.

WATSON, M. W. (1994). Vector autoregressions and cointegration. In *Handbook of Econometrics, Vol. IV. Handbooks in Econom.* **2** 2843–2915. North-Holland, Amsterdam. MR1315982

XU, M. and DUAN, L. L. (2023). Bayesian inference with the l1-ball prior: Solving combinatorial problems with exact zeros. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 1538–1560. MR4718547 https://doi.org/10.1093/jrsssb/qkad076

ZEN, H., TOKUDA, K., MASUKO, T., KOBAYASIH, T. and KITAMURA, T. (2007). A hidden semi-Markov model-based speech synthesis system. *IEICE Trans. Inf. Syst.* **90** 825–834.

ZUCCHINI, W., MACDONALD, I. L. and LANGROCK, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*, 2nd ed. *Monographs on Statistics and Applied Probability* **150**. CRC Press, Boca Raton, FL. MR3618333

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2532–2550
https://doi.org/10.1214/24-AOAS1893

MIXTURE CONDITIONAL REGRESSION WITH ULTRAHIGH DIMENSIONAL TEXT DATA FOR ESTIMATING EXTRALEGAL FACTOR EFFECTS

BY JIAXIN SHI[1,a], FANG WANG[2,d], YUAN GAO[1,b], XIAOJUN SONG[3,e] AND HANSHENG WANG[1,c]

[1]*Guanghua School of Management, Peking University,* [a]*jxshi@stu.pku.edu.cn,* [b]*ygao_stat@outlook.com,* [c]*hansheng@gsm.pku.edu.cn*

[2]*Data Science Institute, Shandong University,* [d]*wangfang226@sdu.edu.cn*

[3]*Guanghua School of Management and Center for Statistical Science, Peking University,* [e]*sxj@gsm.pku.edu.cn*

Testing judicial impartiality is a problem of fundamental importance in empirical legal studies for which standard regression methods have been popularly used to estimate the extralegal factor effects. However, those methods cannot handle control variables with ultrahigh dimensionality, such as those found in judgment documents recorded in text format. To solve this problem, we develop a novel mixture conditional regression (MCR) approach, assuming that the whole sample can be classified into a number of latent classes. Within each latent class, a standard linear regression model can be used to model the relationship between the response and a key feature vector, which is assumed to be of a fixed dimension. Meanwhile, ultrahigh dimensional control variables are then used to determine the latent class membership, where a naïve Bayes type model is used to describe the relationship. Hence, the dimension of control variables is allowed to be arbitrarily high. A novel expectation-maximization algorithm is developed for model estimation. Therefore, we are able to estimate the key parameters of interest as efficiently as if the true class membership were known in advance. Simulation studies are presented to demonstrate the proposed MCR method. A real dataset of Chinese burglary offenses is analyzed for illustration purposes.

REFERENCES

BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.* **45** 77–120. MR3611487 https://doi.org/10.1214/16-AOS1435

BIELEN, S. and GRAJZL, P. (2021). Prosecution or persecution? Extraneous events and prosecutorial decisions. *J. Empir. Leg. Stud.* **18** 765–800.

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.

BILMES, J. A. et al. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *TR—Int. Comput. Sci. Inst.* **4** 126.

BREUSCH, T. S. and PAGAN, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Rev. Econ. Stud.* **47** 239–253. MR0611118 https://doi.org/10.2307/2297111

BRIGHT, S. B. (2008). The failure to achieve fairness: Race and poverty continue to influence who dies. *Univ. Pa. J. Const. Law* **11** 23.

BUSHWAY, S. D. and PIEHL, A. M. (2001). Judging judicial discretion: Legal factors and racial discrimination in sentencing. *Law Soc. Rev.* 733–764.

CANES-WRONE, B., CLARK, T. S. and KELLY, J. P. (2014). Judicial selection and death penalty decisions. *Amer. Polit. Sci. Rev.* **108** 23–39.

DE VEAUX, R. D. (1989). Mixtures of linear regressions. *Comput. Statist. Data Anal.* **8** 227–245. MR1028403 https://doi.org/10.1016/0167-9473(89)90043-1

Key words and phrases. Expectation-maximization algorithm, judicial impartiality, mixture conditional regression, naïve Bayes model, ultrahigh dimensional data.

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

EDMOND, G. (2002). Constructing miscarriages of justice: Misunderstanding scientific evidence in high profile criminal appeals. *Oxf. J. Leg. Stud.* **22** 53–89.

ENGLE, R. F. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handb. Econom.* **2** 775–826.

FELSON, M. and BOBA, R. L. (2010). Crime and everyday life. Sage.

GLYNN, A. N. and SEN, M. (2015). Identifying judicial empathy: Does having daughters cause judges to rule for women's issues? *Amer. J. Polit. Sci.* **59** 37–54.

GROSS, S. and SHAFFER, M. (2012). Exonerations in the United States.

GRÜN, B. and LEISCH, F. (2008). Finite mixtures of generalized linear regression models. In *Recent Advances in Linear Models and Related Areas* 205–230. Springer, Heidelberg. MR2523852 https://doi.org/10.1007/978-3-7908-2064-5_11

JANSEN, R. (1993). Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics* 227–231.

KIM, S.-B., HAN, K.-S., RIM, H.-C. and MYAENG, S. H. (2006). Some effective techniques for naive Bayes text classification. *IEEE Trans. Knowl. Data Eng.* **18** 1457–1466.

KONONENKO, I. (1991). Semi-naive Bayesian classifier. In *Machine Learning—EWSL*-91 (*Porto*, 1991). *Lecture Notes in Computer Science* **482** 206–219. Springer, Berlin. MR1101397 https://doi.org/10.1007/BFb0017015

KREHBIEL, P. J. and CROPANZANO, R. (2000). Procedural justice, outcome favorability and emotion. *Soc. Justice Res.* **13**.

L'HEUREUX-DUBE, C. (2001). Beyond the myths: Qquality, impartiality, and justice. *J. Soc. Distress Homeless*.

LYNCH, M. and HANEY, C. (2011). Mapping the racial bias of the white male capital juror: Jury composition and the "Empathic divide". *Law Soc. Rev.* **45** 69–102.

MERON, T. (2005). Judicial independence and impartiality in international criminal tribunals. *Amer. J. Int. Law* **99** 359–369.

MEYERSON, D. (2006). *Understanding Jurisprudence*, 1st ed. Routledge, London.

MINNIER, J., YUAN, M., LIU, J. S. and CAI, T. (2015). Risk classification with an adaptive naive Bayes kernel machine model. *J. Amer. Statist. Assoc.* **110** 393–404. MR3338511 https://doi.org/10.1080/01621459.2014.908778

MISHLER, W. and SHEEHAN, R. S. (1993). The Supreme Court as a countermajoritarian institution? The impact of public opinion on Supreme Court decisions. *Amer. Polit. Sci. Rev.* **87** 87–101.

NOBLES, R. and SCHIFF, D. (1995). Miscarriages of justice: A systems approach. *Mod. Law Rev.* **58** 299.

PENG, Y. and CHENG, J. (2022). Ethnic disparity in Chinese theft sentencing. *China Rev.* **22** 47–71.

QAISER, S. and ALI, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. *Int. J. Comput. Appl.* **181** 25–29.

RAMOS, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning* **242** 29–48, NJ, USA.

REYNOLDS, D. A. (2009). Gaussian mixture models. *Encycl. Biometrics* **741**.

RISH, I. et al. (2001). An empirical study of the naive Bayes classifier. In *IJCAI* 2001 *Workshop on Empirical Methods in Artificial Intelligence* **3** 41–46.

ROBERTS, S. (2003). 'Unsafe' convictions: Defining and compensating miscarriages of justice. *Mod. Law Rev.* **66** 441–451.

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

SEDGHI, H., JANZAMIN, M. and ANANDKUMAR, A. (2016). Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics* 1223–1231. PMLR.

SHAO, J. (2003). *Mathematical Statistics*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR2002723 https://doi.org/10.1007/b97553

SHELEY, J. F. and ASHKINS, C. D. (1981). Crime, crime news, and crime views. *Public Opin. Q.* **45** 492–506.

SHI, J., WANG, F., GAO, Y., SONG, X. and WANG, H. (2024). Supplement to "Mixture conditional regression with ultrahigh dimensional text data for estimating extralegal factor effects." https://doi.org/10.1214/24-AOAS1893SUPPA, https://doi.org/10.1214/24-AOAS1893SUPPB

SIMMONS, J. and FLOOD-PAGE, C. (2002). *Crime in England and Wales*. Home Office, London.

SPIEGELHALTER, D. J. and KNILL-JONES, R. P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *J. R. Stat. Soc., A* **147** 35–58.

STEFFENSMEIER, D. and KRAMER, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology* **36** 763–798.

STITH, K., CABRANES, J. A. et al. (1998). *Fear of Judging*: *Sentencing Guidelines in the Federal Courts*. Univ. Chicago Press, Chicago.

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

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

VERMUNT, J. K. and MAGIDSON, J. (2002). Latent class cluster analysis. In *Applied Latent Class Analysis* 89–106. Cambridge Univ. Press, Cambridge. MR1927666 https://doi.org/10.1017/CBO9780511499531.004

WADHAM, J. (1993). Unravelling miscarriages of justice. *New Law J*. **143** 1650–1650.

WEDEL, M., KAMAKURA, W. A., WEDEL, M. and KAMAKURA, W. A. (2000). *Mixture Regression Models*. Springer, Berlin.

WEILER, P. (1968). Two models of judicial decision-making. *Canadian Bar Review* **46** 406.

WU, H. C., LUK, R. W. P., WONG, K. F. and KWOK, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. *ACM Trans*. *Inf*. *Syst*. **26** 1–37.

XU, K., LIU, H., WANG, F. and WANG, H. (2022). 'This crime is not that rrime'-classification and evaluation of four common crimes. *Law Probab. Risk* **20** 135–152.

XU, L. and JORDAN, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Comput*. **8** 129–151.

YANG, F.-J. (2018). An implementation of naive Bayes classifier. In 2018 *International Conference on Computational Science and Computational Intelligence* (*CSCI*) 301–306. IEEE Press, New York.

YE, X. (2010). The impact and direction of national standardized sentencing reform in China. *Columbia J*. *Asian Law* **24** 247.

ZHAO, J., JIN, L. and SHI, L. (2015). Mixture model selection via hierarchical BIC. *Comput*. *Statist*. *Data Anal*. **88** 139–153. MR3332023 https://doi.org/10.1016/j.csda.2015.01.019

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2551–2575
https://doi.org/10.1214/24-AOAS1894
© Institute of Mathematical Statistics, 2024

A FLEXIBLE MODEL FOR CORRELATED COUNT DATA, WITH APPLICATION TO MULTICONDITION DIFFERENTIAL EXPRESSION ANALYSES OF SINGLE-CELL RNA SEQUENCING DATA

BY YUSHA LIU[1,a], PETER CARBONETTO[2,b], MICHIHIRO TAKAHAMA[3,c], ADAM GRUENBAUM[4,e], DONGYUE XIE[5,f], NICOLAS CHEVRIER[3,d], AND MATTHEW STEPHENS[6,g]

[1]*Department of Biostatistics, University of North Carolina at Chapel Hill,* [a]*yushaliu@unc.edu*
[2]*Department of Human Genetics, University of Chicago,* [b]*pcarbo@uchicago.edu*
[3]*Pritzker School of Molecular Engineering, University of Chicago,* [c]*mtakahama@uchicago.edu,* [d]*nchevrier@uchicago.edu*
[4]*Institute for Health Metrics and Evaluation, University of Washington,* [e]*adam.gruenbaum@gmail.com*
[5]*Department of Statistics, University of Chicago,* [f]*dyxie@uchicago.edu*
[6]*Departments of Statistics and Human Genetics, University of Chicago,* [g]*mstephens@uchicago.edu*

Detecting differences in gene expression is an important part of single-cell RNA sequencing experiments, and many statistical methods have been developed for this aim. Most differential expression analyses focus on comparing expression between two groups (e.g., treatment vs. control). But there is increasing interest in *multicondition differential expression analyses* in which expression is measured in many conditions and the aim is to accurately detect and estimate expression differences in all conditions. We show that directly modeling single-cell RNA-seq counts in all conditions simultaneously, while also inferring how expression differences are shared across conditions, leads to greatly improved performance for detecting and estimating expression differences compared to existing methods. We illustrate the potential of this new approach by analyzing data from a single-cell experiment studying the effects of cytokine stimulation on gene expression. We call our new method "Poisson multivariate adaptive shrinkage," and it is implemented in an R package available at https://github.com/stephenslab/poisson.mash.alpha.

REFERENCES

AHLMANN-ELTZE, C. and HUBER, W. (2020). glmGamPoi: Fitting Gamma–Poisson generalized linear models on single cell count data. *Bioinformatics* **36** 5701–5702.

AITCHISON, J. and HO, C.-H. (1989). The multivariate Poisson-log normal distribution. *Biometrika* **76** 643–653. MR1041409 https://doi.org/10.1093/biomet/76.4.643

ALTMEIER, S., TOSKA, A., SPARBER, F., TEIJEIRA, A., HALIN, C. and LEIBUNDGUT-LANDMANN, S. (2016). IL-1 coordinates the neutrophil response to C. albicans in the oral mucosa. *PLoS Pathog*. **12** e1005882.

ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*. **11** R106.

ARRIDGE, S. R., ITO, K., JIN, B. and ZHANG, C. (2018). Variational Gaussian approximation for Poisson data. *Inverse Probl*. **34** 025005, 29. MR3751049 https://doi.org/10.1088/1361-6420/aaa0ab

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

BOCHKINA, N. and RICHARDSON, S. (2007). Tail posterior probability for inference in pairwise and multiclass gene expression data. *Biometrics* **63** 1117–1125, 1312. MR2414589 https://doi.org/10.1111/j.1541-0420.2007.00807.x

BOOTHBY, I. C., COHEN, J. N. and ROSENBLUM, M. D. (2020). Regulatory T cells in skin injury: At the crossroads of tolerance and tissue repair. *Sci. Immunol*. **5** eaaz9631.

BULLARD, J. H., PURDOM, E., HANSEN, K. D. and DUDOIT, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinform*. **11** 94.

Key words and phrases. Poisson, multivariate count data, empirical Bayes, variational inference, single-cell RNA sequencing, differential expression analysis.

Chang, J., Burkett, P. R., Borges, C. M., Kuchroo, V. K., Turka, L. A. and Chang, C.-H. (2013). MyD88 is essential to sustain mTOR activation necessary to promote T helper 17 cell proliferation by linking IL-1 and IL-23 signaling. *Proc. Natl. Acad. Sci. USA* **110** 2270–2275.

Cooper, A. M., Magram, J., Ferrante, J. and Orme, I. M. (1997). Interleukin 12 (IL-12) is crucial to the development of protective immunity in mice intravenously infected with mycobacterium tuberculosis. *J. Exp. Med.* **186** 39–45.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley, Hoboken, NJ. MR2239987

Crowell, H. L., Soneson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D. and Robinson, M. D. (2020). Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11** 6077.

Cruz, A., Khader, S. A., Torrado, E., Fraga, A., Pearl, J. E., Pedrosa, J., Cooper, A. M. and Castro, A. G. (2006). Cutting edge: IFN-γ regulates the induction and expansion of IL-17-producing CD4 T cells during mycobacterial infection. *J. Immunol.* **177** 1416–1420.

Dinarello, C. A. (2018). Overview of the IL-1 family in innate inflammation and acquired immunity. *Immunol. Rev.* **281** 8–27.

Dinarello, C. A., Simon, A. and Van Der Meer, J. W. (2012). Treating inflammation by blocking interleukin-1 in a broad spectrum of diseases. *Nat. Rev. Drug Discov.* **11** 633–652.

Dreis, C., Ottenlinger, F. M., Putyrski, M., Ernst, A., Huhn, M., Schmidt, K. G., Pfeilschifter, J. M. and Radeke, H. H. (2019). Tissue cytokine IL-33 modulates the cytotoxic CD8 T lymphocyte activity during nutrient deprivation by regulation of lineage-specific differentiation programs. *Front. Immunol.* **1698**.

Erdmann-Pham, D. D., Fischer, J., Hong, J. and Song, Y. S. (2021). Likelihood-based deconvolution of bulk gene expression data using single-cell references. *Genome Res.* **31** 1794–1806.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K. et al. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16** 278.

Gao, L. L., Bien, J. and Witten, D. (2024). Selective inference for hierarchical clustering. *J. Amer. Statist. Assoc.* **119** 332–342. MR4713896 https://doi.org/10.1080/01621459.2022.2116331

Gerard, D. (2020). Data-based RNA-seq simulations by binomial thinning. *BMC Bioinform.* **21** 206.

Gerard, D. and Stephens, M. (2020). Empirical Bayes shrinkage and false discovery rate estimation, allowing for unwanted variation. *Biostatistics* **21** 15–32. MR4043843 https://doi.org/10.1093/biostatistics/kxy029

Gu, J., Wang, X., Halakivi-Clarke, L., Clarke, R. and Xuan, J. (2014). BADGE: A novel Bayesian model for accurate abundance quantification and differential analysis of RNA-seq data. *BMC Bioinform.* **15** S6.

Jabri, B. and Abadie, V. (2015). IL-15 functions as a danger signal to regulate tissue-resident T cells and tissue destruction. *Nat. Rev., Immunol.* **15** 771–783.

Kang, G., Du, L. and Zhang, H. (2016). MultiDE: A dimension reduced model based statistical method for differential expression analysis using RNA-sequencing data with multiple treatment conditions. *BMC Bioinform.* **17** 248.

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

Kuhn, J. A., Vainchtein, I. D., Braz, J., Hamel, K., Bernstein, M., Craik, V. et al. (2021). Regulatory T-cells inhibit microglia-induced pain hypersensitivity in female mice. *eLife* **10** e69056.

Law, C. W., Chen, Y., Shi, W. and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15** R29.

Leek, J. T. (2014). svaseq: Removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42** e161.

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

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. and Zhang, B. (2019). WebGestalt 2019: Gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47** W199–W205.

Littman, D. R. and Rudensky, A. Y. (2010). Th17 and regulatory T cells in mediating and restraining inflammation. *Cell* **140** 845–858.

Liu, Y., Carbonetto, P., Takahama, M., Gruenbaum, A., Xie, D., Chevrier, N. and Stephens, M. (2024). Supplement to “A flexible model for correlated count data, with application to multicondition differential expression analyses of single-cell RNA sequencing data.” https://doi.org/10.1214/24-AOAS1894SUPPA, https://doi.org/10.1214/24-AOAS1894SUPPB, https://doi.org/10.1214/24-AOAS1894SUPPC

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.

LUN, A. (2018). Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data. bioRxiv. https://doi.org/10.1101/404962

LUN, A. T. L. and MARIONI, J. C. (2017). Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics* **18** 451–464. MR3799589 https://doi.org/10.1093/biostatistics/kxw055

MCCARTHY, D. J. and SMYTH, G. K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* **25** 765–771.

MCGEACHY, M. J., CHEN, Y., TATO, C. M., LAURENCE, A., JOYCE-SHAIKH, B., BLUMENSCHEIN, W. M., MCCLANAHAN, T. K., O'SHEA, J. J. and CUA, D. J. (2009). The interleukin 23 receptor is essential for the terminal differentiation of interleukin 17–producing effector T helper cells in vivo. *Nat. Immunol.* **10** 314–324.

MURPHY, A. E. and SKENE, N. G. (2022). A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. *Nat. Commun.* **13** 7851.

OKAMURA, H., TSUTSUI, H., KOMATSU, T., YUTSUDO, M., HAKURA, A., TANIMOTO, T. et al. (1995). Cloning of a new cytokine that induces IFN-γ production by T cells. *Nature* **378** 88–91.

RISSO, D., NGAI, J., SPEED, T. P. and DUDOIT, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32** 896–902.

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.

ROBINSON, M. D. and SMYTH, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9** 321–332.

SARKAR, A. and STEPHENS, M. (2021). Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet.* **53** 770–777.

SHIMOBAYASHI, M. and HALL, M. N. (2014). Making new contacts: The mTOR network in metabolism and signalling crosstalk. *Nat. Rev., Mol. Cell Biol.* **15** 155–162.

SILVA, A., ROTHSTEIN, S. J., MCNICHOLAS, P. D. and SUBEDI, S. (2019). A multivariate Poisson-log normal mixture model for clustering transcriptome sequencing data. *BMC Bioinform.* **20** 394.

SMYTH, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3** Art. 3, 29. MR2101454 https://doi.org/10.2202/1544-6115.1027

SONESON, C. and DELORENZI, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform.* **14** 91.

SONESON, C. and ROBINSON, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15** 255–261.

SQUAIR, J. W., GAUTIER, M., KATHE, C., ANDERSON, M. A., JAMES, N. D., HUTSON, T. H. et al. (2021). Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12** 5692.

STEPHENS, M. (2017). False discovery rates: A new deal. *Biostatistics* **18** 275–294. MR3824755 https://doi.org/10.1093/biostatistics/kxw041

SUBEDI, S. and BROWNE, R. P. (2020). A family of parsimonious mixtures of multivariate Poisson-lognormal distributions for clustering multivariate count data. *Stat* **9** e310, 11. MR4193415 https://doi.org/10.1002/sta4.310

THE GENE ONTOLOGY CONSORTIUM (2020). The gene ontology resource: Enriching a GOld mine. *Nucleic Acids Res.* **49** D325–D334.

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** 295.

URBUT, S. M., WANG, G., CARBONETTO, P. and STEPHENS, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51** 187–195.

WANG, T., LI, B., NELSON, C. E. and NABAVI, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinform.* **20** 40.

WANG, W. and STEPHENS, M. (2021). Empirical Bayes matrix factorization. *J. Mach. Learn. Res.* **22** Paper No. 120, 1–40. MR4279771 https://doi.org/10.1007/s00023-020-00971-9

WANG, Z., GERSTEIN, M. and SNYDER, M. (2009). RNA-seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10** 57–63.

WEI, Y., TENZEN, T. and JI, H. (2015). Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics* **16** 31–46. MR3365409 https://doi.org/10.1093/biostatistics/kxu038

WOJNO, E. D. T., HUNTER, C. A. and STUMHOFER, J. S. (2019). The immunobiology of the interleukin-12 family: Room for discovery. *Immunity* **50** 851–870.

YOSHIMOTO, T., OKAMURA, H., TAGAWA, Y.-I., IWAKURA, Y. and NAKANISHI, K. (1997). Interleukin 18 together with interleukin 12 inhibits IgE production by induction of interferon-γ production from activated B cells. *Proc. Natl. Acad. Sci. USA* **94** 3948–3953.

ZHANG, M., LIU, S., MIAO, Z., HAN, F., GOTTARDO, R. and SUN, W. (2022). IDEAS: Individual level differential expression analysis for single-cell RNA-seq data. *Genome Biol.* **23** 33.

ZHENG, G. X., TERRY, J. M., BELGRADER, P., RYVKIN, P., BENT, Z. W., WILSON, R. et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8** 14049.

ZHU, A., IBRAHIM, J. G. and LOVE, M. I. (2019). Heavy-tailed prior distributions for sequence count data: Removing the noise and preserving large differences. *Bioinformatics* **35** 2084–2092.

ZHU, J., GUO, L., MIN, B., WATSON, C. J., HU-LI, J., YOUNG, H. A., TSICHLIS, P. N. and PAUL, W. E. (2002). Growth factor independent-1 induced by IL-4 regulates Th2 cell proliferation. *Immunity* **16** 733–744.

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2576–2595
https://doi.org/10.1214/24-AOAS1896
© Institute of Mathematical Statistics, 2024

WEIGHTED BIOMARKER VARIABILITY IN JOINT ANALYSIS OF LONGITUDINAL AND TIME-TO-EVENT DATA

BY CHUNYU WANG[1,a], JIAMING SHEN[2,b], CHRISTIANA CHARALAMBOUS[2,c] AND JIANXIN PAN[3,d]

[1]*MRC Biostatistics Unit, University of Cambridge,* [a]*chunyu.wang@mrc-bsu.cam.ac.uk*

[2]*Department of Mathematics, The University of Manchester,* [b]*jiaming.shen@manchester.ac.uk,* [c]*christiana.charalambous@manchester.ac.uk*

[3]*Research Center for Mathematics, Beijing Normal University and Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College,* [d]*jianxinpan@uic.edu.cn*

Motivated by the clinical evidence that the biomarker variability may have prognostic value for a related disease, we extend the standard joint model for longitudinal and time-to-event outcomes to incorporate the weighted cumulative effects of both biomarker level and variability on the survival hazard. A mixed-effects model is specified for biomarker observations wherein the subject-specific trajectories are modelled by spline functions with random coefficients. Borrowing ideas from smoothing splines, we propose a new variability measure which characterizes the roughness of the subject-specific biomarker trajectory by the integrated amount of its second derivatives over time. The inclusion of weight functions in cumulative quantities permits the importance of biomarker history to vary with time. To reduce computational complexity, we confine the weight functions to a particular parametric family with scale parameters to be estimated. Asymptotic properties of maximum likelihood estimators are established with a discussion on the identification issue of the scale parameters. We use EM algorithm in estimation with initial values obtained from a two-stage method. Simulation studies have been conducted under different settings. Finally, we apply our model to investigate the weighted cumulative effects of systolic blood pressure level and variability on cardiovascular events in the Medical Research Council trial.

REFERENCES

ABRAHAMOWICZ, M., BARTLETT, G., TAMBLYN, R. and DU BERGER, R. (2006). Modeling cumulative dose and exposure duration provided insights regarding the associations between benzodiazepines and injuries. *J. Clin. Epidemiol.* **59** 393–403.

BARRETT, J. K., HUILLE, R., PARKER, R., YANO, Y. and GRISWOLD, M. (2019). Estimating the association between blood pressure variability and cardiovascular disease: An application using the ARIC study. *Stat. Med.* **38** 1855–1868. MR3934823 https://doi.org/10.1002/sim.8074

BRESLOW, N. E., LUBIN, J., MAREK, P. and LANGHOLZ, B. (1983). Multiplicative models and cohort analysis. *J. Amer. Statist. Assoc.* **78** 1–12.

CARR, M. J., BAO, Y., PAN, J., CRUICKSHANK, K. and MCNAMEE, R. (2012). The predictive ability of blood pressure in elderly trial patients. *J. Hypertens.* **30** 1725–1733. https://doi.org/10.1097/HJH.0b013e3283568a73

DANIELI, C., SHEPPARD, T., COSTELLO, R., DIXON, W. G. and ABRAHAMOWICZ, M. (2020). Modeling of cumulative effects of time-varying drug exposures on within-subject changes in a continuous outcome. *Stat. Methods Med. Res.* **29** 2554–2568. MR4129429 https://doi.org/10.1177/0962280220902179

GAO, F., MILLER, J. P., XIONG, C., BEISER, J. A., GORDON, M. and GROUP, T. O. H. T. S. (2011). A joint-modeling approach to assess the impact of biomarker variability on the risk of developing clinical outcome. *Stat. Methods Appl.* **20** 83–100. MR2771020 https://doi.org/10.1007/s10260-010-0150-z

HAUPTMANN, M., WELLMANN, J., LUBIN, J. H., ROSENBERG, P. S. and KREIENBROCK, L. (2000). Analysis of exposure-time–response relationships using a spline weight function. *Biometrics* **56** 1105–1108. MR1815589 https://doi.org/10.1111/j.0006-341X.2000.01105.x

Key words and phrases. Variability, weighted cumulative effect, joint modelling, spline, MRC trial.

HOWARD, S. C. and ROTHWELL, P. M. (2009). Reproducibility of measures of visit-to-visit variability in blood pressure after transient ischaemic attack or minor stroke. *Cerebrovasc. Dis.* **28** 331–340. https://doi.org/10.1159/000229551

JOHANSEN, S. (1983). An extension of Cox's regression model. *Int. Stat. Rev.* **51** 165–174. MR0715533 https://doi.org/10.2307/1402746

LEVER, A. F. and BRENNAN, P. J. (1993). MRC trial of treatment in elderly hypertensives. *Clin. Exp. Hypertens.* **15** 941–952. https://doi.org/10.3109/10641969309037083

LEWBEL, A. (2019). The identification zoo: Meanings of identification in econometrics. *J. Econ. Lit.* **57** 835–903.

LI, X. (2019). Modeling exposure-time-response association in the presence of competing risks. PhD thesis, Univ. Pittsburgh. MR4024247

LYLES, R. H., MUNÕZ, A., XU, J., TAYLOR, J. M. and CHMIEL, J. S. (1999). Adjusting for measurement error to assess health effects of variability in biomarkers. *Stat. Med.* **18** 1069–1086.

MAUFF, K., STEYERBERG, E. W., NIJPELS, G., VAN DER HEIJDEN, A. A. W. A. and RIZOPOULOS, D. (2017). Extension of the association structure in joint models to include weighted cumulative effects. *Stat. Med.* **36** 3746–3759. MR3696554 https://doi.org/10.1002/sim.7385

MUNTNER, P., JOYCE, C., LEVITAN, E. B., HOLT, E., SHIMBO, D., WEBBER, L. S., OPARIL, S., RE, R. and KROUSEL-WOOD, M. (2011). Reproducibility of visit-to-visit variability of blood pressure measured as part of routine clinical care. *J. Hypertens.* **29** 2332–2338. https://doi.org/10.1097/HJH.0b013e32834cf213

PARTY, M. W. et al. (1992). Medical research council trial of treatment of hypertension in older adults: Principal results. *BMJ* **304** 405–412.

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

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

ROTHWELL, P. M. (2010). Limitations of the usual blood-pressure hypothesis and importance of variability, instability, and episodic hypertension. *Lancet* **375** 938–948.

ROTHWELL, P. M., HOWARD, S. C., DOLAN, E., O'BRIEN, E., DOBSON, J. E., DAHLÖF, B., POULTER, N. R., SEVER, P. S. et al. (2010a). Effects of β blockers and calcium-channel blockers on within-individual variability in blood pressure and risk of stroke. *Lancet Neurol.* **9** 469–480.

ROTHWELL, P. M., HOWARD, S. C., DOLAN, E., O'BRIEN, E., DOBSON, J. E., DAHLÖF, B., SEVER, P. S. and POULTER, N. R. (2010b). Prognostic significance of visit-to-visit variability, maximum systolic blood pressure, and episodic hypertension. *Lancet* **375** 895–905.

SHI, M., WEISS, R. E. and TAYLOR, J. M. (1996). An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **45** 151–163.

SYLVESTRE, M.-P. and ABRAHAMOWICZ, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Stat. Med.* **28** 3437–3453. MR2744373 https://doi.org/10.1002/sim.3701

TIERNEY, L., KASS, R. E. and KADANE, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.* **84** 710–716. MR1132586

VACEK, P. M. (1997). Assessing the effect of intensity when exposure varies over time. *Stat. Med.* **16** 505–513. https://doi.org/10.1002/(sici)1097-0258(19970315)16:5<505::aid-sim424>3.0.co;2-z

WANG, C., SHEN, J., CHARALAMBOUS, C. and PAN, J. (2024a). Modeling biomarker variability in joint analysis of longitudinal and time-to-event data. *Biostatistics* **25** 577–596. MR4732249 https://doi.org/10.1093/biostatistics/kxad009

WANG, C., SHEN, J., CHARALAMBOUS, C. and PAN, J. (2024). Supplement to "Weighted biomarker variability in joint analysis of longitudinal and time-to-event data." https://doi.org/10.1214/24-AOAS1896SUPPA, https://doi.org/10.1214/24-AOAS1896SUPPB

YANO, Y. (2017). Visit-to-visit blood pressure variability-what is the current challenge? *Am. J. Hypertens.* **30** 112–114. https://doi.org/10.1093/ajh/hpw124

YU, D., PEAT, G., BEDSON, J., EDWARDS, J. J., TURKIEWICZ, A. and JORDAN, K. P. (2016). Weighted cumulative exposure models helped identify an association between early knee-pain consultations and future knee OA diagnosis. *J. Clin. Epidemiol.* **76** 218–228. https://doi.org/10.1016/j.jclinepi.2016.02.025

ZENG, D. and CAI, J. (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Ann. Statist.* **33** 2132–2163. MR2211082 https://doi.org/10.1214/009053605000000480

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2596–2614
https://doi.org/10.1214/24-AOAS1897

BIVARIATE FUNCTIONAL PATTERNS OF LIFETIME MEDICARE COSTS AMONG ESRD PATIENTS

BY YUE WANG[1,a], BIN NAN[1,b] AND JOHN D. KALBFLEISCH[2,c]

[1]*Department of Statistics, University of California, Irvine,* [a]*ywang47@uci.edu,* [b]*nanb@uci.edu*
[2]*Department of Biostatistics, University of Michigan,* [c]*jdkalbfl@umich.edu*

In this work we study the lifetime Medicare spending patterns of patients with end-stage renal disease (ESRD). We extract the information of patients who started their ESRD services in 2007–2011 from the United States Renal Data System (USRDS). Patients are partitioned into three groups based on their kidney transplant status: 1-unwaitlisted and never transplanted, 2-waitlisted but never transplanted, and 3-waitlisted and then transplanted. To study their Medicare cost trajectories, we use a semiparametric regression model with both fixed and bivariate time-varying coefficients to compare groups 1 and 2 as well as a bivariate time-varying coefficient model with different starting times (time since the first ESRD service and time since the kidney transplant) to compare groups 2 and 3. In addition to demographics and other medical conditions, these regression models are conditional on the survival time, which ideally depict the lifetime Medicare spending patterns. For estimation we extend the profile weighted least squares (PWLS) estimator to longitudinal data for the first comparison and propose a two-stage estimating method for the second comparison. We use sandwich variance estimators to construct confidence intervals and validate inference procedures through simulations. Our analysis of the Medicare claims data reveals that waitlisting is associated with a lower daily medical cost at the beginning of ESRD service among waitlisted patients, which gradually increases over time. Averaging over lifespan, however, there is no difference between waitlisted and unwaitlisted groups. A kidney transplant, on the other hand, reduces the medical cost significantly after an initial spike.

REFERENCES

BERGER, A., EDELSBERG, J., INGLESE, G. W., BHATTACHARYYA, S. K. and OSTER, G. (2009). Cost comparison of peritoneal dialysis versus hemodialysis in end-stage renal disease. *Amer. J. Manag. Care* **15** 509–518.

CAMERON, D., UBELS, J. and NORSTRÖM, F. (2018). On what basis are medical cost-effectiveness thresholds set? Clashing opinions and an absence of data: A systematic review. *Glob. Health Action* **11** 1447828.

DRUMMOND, M. F., SCULPHER, M. J., CLAXTON, K., STODDART, G. L. and TORRANCE, G. W. (2015). *Methods for the Economic Evaluation of Health Care Programmes*. Oxford University Press, London.

FAN, J., HUANG, T. and LI, R. (2007). Analysis of longitudinal data with semiparametric estimation of convariance function. *J. Amer. Statist. Assoc.* **102** 632–641. MR2370857 https://doi.org/10.1198/016214507000000095

FU, R., SEKERCIOGLU, N., BERTA, W. and COYTE, P. C. (2020). Cost-effectiveness of deceased-donor renal transplant versus dialysis to treat end-stage renal disease: A systematic review. *Transplant. Direct* **6** e522. https://doi.org/10.1097/TXD.0000000000000974

KONG, S., NAN, B., KALBFLEISCH, J. D., SARAN, R. and HIRTH, R. (2018). Conditional modeling of longitudinal data with terminal event. *J. Amer. Statist. Assoc.* **113** 357–368. MR3803470 https://doi.org/10.1080/01621459.2016.1255637

LIU, L., WOLFE, R. A. and KALBFLEISCH, J. D. (2007). A shared random effects model for censored medical costs and mortality. *Stat. Med.* **26** 139–155. MR2312704 https://doi.org/10.1002/sim.2535

MACHNICKI, G., SERIAI, L. and SCHNITZLER, M. A. (2006). Economics of transplantation: A review of the literature. *Transplant. Rev.* **20** 61–75.

Key words and phrases. Bivariate varying-coefficient model, end-stage renal disease, kernel smoothing, kidney transplant, waiting list.

MANDELBROT, D. A., FLEISHMAN, A., RODRIGUE, J. R., NORMAN, S. P. and SAMANIEGO, M. (2017). Practices in the evaluation of potential kidney transplant recipients who are elderly: A survey of US transplant centers. *Clin. Transplant.* **31** e13088.

MASSIE, A., LUO, X., CHOW, E., ALEJO, J., DESAI, N. and SEGEV, D. (2014). Survival benefit of primary deceased donor transplantation with high-KDPI kidneys. *Am. J. Transplant.* **14** 2310–2316.

MCADAMS-DEMARCO, M. A., RASMUSSEN, S. E. V. P., CHU, N. M., AGOONS, D., PARSONS, R. F., AL-HAMAD, T., JOHANSEN, K. L., TULLIUS, S. G., LYNCH, R. et al. (2020). Perceptions and practices regarding frailty in kidney transplantation: Results of a national survey. *Transplantation* **104** 349.

MENZIN, J., LINES, L. M., WEINER, D. E., NEUMANN, P. J., NICHOLS, C., RODRIGUEZ, L., AGODOA, I. and MAYNE, T. (2011). A review of the costs and cost effectiveness of interventions in chronic kidney disease: Implications for policy. *PharmacoEconomics* **29** 839–861. https://doi.org/10.2165/11588390-000000000-00000

SHEN, S.-L., CUI, J.-L., MEI, C.-L. and WANG, C.-W. (2014). Estimation and inference of semi-varying coefficient models with heteroscedastic errors. *J. Multivariate Anal.* **124** 70–93. MR3147312 https://doi.org/10.1016/j.jmva.2013.10.010

SUN, Y., QI, L., HENG, F. and GILBERT, P. B. (2019). Analysis of generalized semiparametric mixed varying-coefficients models for longitudinal data. *Canad. J. Statist.* **47** 352–373. MR3997994 https://doi.org/10.1002/cjs.11498

UNITED STATES RENAL DATA SYSTEM (2019). Researcher's guide to the USRDS database. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD.

UNITED STATES RENAL DATA SYSTEM (2022). USRDS annual data report: Epidemiology of kidney disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD.

TONELLI, M., WIEBE, N., KNOLL, G., BELLO, A., BROWNE, S., JADHAV, D., KLARENBACH, S. and GILL, J. (2011). Systematic review: Kidney transplantation compared with dialysis in clinically relevant outcomes. *Am. J. Transplant.* **11** 2093–2109. https://doi.org/10.1111/j.1600-6143.2011.03686.x

WANG, Y., NAN, B. and KALBFLEISCH, J. D. (2023). Kernel estimation of bivariate time-varying coefficient model for longitudinal data with terminal event. *J. Amer. Statist. Assoc.* In press.

WANG, Y., NAN, B. and KALBFLEISCH, J. D (2024). Supplement to "Bivariate Functional Patterns of Lifetime Medicare Costs among ESRD Patients." https://doi.org/10.1214/24-AOAS1897SUPPA, https://doi.org/10.1214/24-AOAS1897SUPPB, https://doi.org/10.1214/24-AOAS1897SUPPC

WHELAN, A. M., JOHANSEN, K. L., COPELAND, T., MCCULLOCH, C. E., NALLAPOTHULA, D., LEE, B. K., ROLL, G. R., WEIR, M. R., ADEY, D. B. et al. (2022). Kidney transplant candidacy evaluation and waitlisting practices in the United States and their association with access to transplantation. *Am. J. Transplant.* **22** 1624–1636.

WOLFE, R. A., ASHBY, V. B., MILFORD, E. L., OJO, A. O., ETTENGER, R. E., AGODOA, L. Y., HELD, P. J. and PORT, F. K. (1999). Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *N. Engl. J. Med.* **341** 1725–1730. https://doi.org/10.1056/NEJM199912023412303

WOLFE, R. A., MCCULLOUGH, K. P. and LEICHTMAN, A. B. (2009). Predictability of survival models for waiting list and transplant patients: Calculating LYFT. *Am. J. Transplant.* **9** 1523–1527. https://doi.org/10.1111/j.1600-6143.2009.02708.x

WU, C. O., CHIANG, C.-T. and HOOVER, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Amer. Statist. Assoc.* **93** 1388–1402. MR1666635 https://doi.org/10.2307/2670054

ZHANG, W., LEE, S.-Y. and SONG, X. (2002). Local polynomial fitting in semivarying coefficient model. *J. Multivariate Anal.* **82** 166–188. MR1918619 https://doi.org/10.1006/jmva.2001.2012

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2615–2634
https://doi.org/10.1214/24-AOAS1899

ARE MADE AND MISSED DIFFERENT? AN ANALYSIS OF FIELD GOAL ATTEMPTS OF PROFESSIONAL BASKETBALL PLAYERS VIA DEPTH BASED TESTING PROCEDURE

BY KAI QI[1,a], GUANYU HU[2,b] AND WEI WU[3,c]

[1]*Microsoft*, [a]*kq15b@my.fsu.edu*

[2]*Center for Spatial Temporal Modeling for Applications in Population Sciences, Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston*, [b]*guanyu.hu@uth.tmc.edu*

[3]*Department of Statistics, Florida State University*, [c]*wwu@fsu.edu*

In this paper we develop a novel depth-based testing procedure on spatial point processes to examine the difference in made and missed field goal attempts for NBA players. Specifically, our testing procedure can statistically detect the differences between made and missed field goal attempts for NBA players. We first obtain the depths of two processes under the polar coordinate system. A two-dimensional Kolmogorov–Smirnov test is then performed to test the difference between the depths of the two processes. Throughout extensive simulation studies, we show our testing procedure with good frequentist properties under both null hypothesis and alternative hypothesis. A comparison against the competing methods shows that our proposed procedure has better testing reliability and testing power. Application to the shot chart data of 191 NBA players in the 2017–2018 regular season offers interesting insights about these players' made and missed shot patterns.

REFERENCES

BADDELEY, A. (2017). Local composite likelihood for spatial point processes. *Spat. Stat.* **22** 261–295. MR3732850 https://doi.org/10.1016/j.spasta.2017.03.001

BADDELEY, A. and TURNER, R. (2005). spatstat: An R package for analyzing spatial point patterns. *J. Stat. Softw.* **12** 1–42.

BARNETT, V. (1976). The ordering of multivariate data. *J. Roy. Statist. Soc. Ser. A* **139** 318–355. MR0445726 https://doi.org/10.2307/2344839

BROWN, B. M., HETTMANSPERGER, T. P., NYBLOM, J. and OJA, H. (1992). On certain bivariate sign tests and medians. *J. Amer. Statist. Assoc.* **87** 127–135. MR1158630

CERVONE, D., D'AMOUR, A., BORNN, L. and GOLDSBERRY, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *J. Amer. Statist. Assoc.* **111** 585–599. MR3538688 https://doi.org/10.1080/01621459.2016.1141685

CHRISTMANN, A. (2002). Classification based on the support vector machine and on regression depth. In *Statistical Data Analysis Based on the L_1-Norm and Related Methods* (*Neuchâtel*, 2002). *Stat. Ind. Technol.* 341–352. Birkhäuser, Basel. MR2001327

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

DONOHO, D. L. and GASKO, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* **20** 1803–1827. MR1193313 https://doi.org/10.1214/aos/1176348890

ERČULJ, F. and ŠTRUMBELJ, E. (2015). Basketball shot types and shot success in different levels of competitive basketball. *PLoS ONE* **10** e0128885.

FASANO, G. and FRANCESCHINI, A. (1987). A multidimensional version of the Kolmogorov–Smirnov test. *Mon. Not. R. Astron. Soc.* **225** 155–170.

FRAIMAN, R., LIU, R. Y. and MELOCHE, J. (1997). Multivariate density estimation by probing depth. In *L_1-Statistical Procedures and Related Topics* (*Neuchâtel*, 1997). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **31** 415–430. IMS, Hayward, CA. MR1833602 https://doi.org/10.1214/lnms/1215454155

Key words and phrases. Basketball shot charts, hypothesis testing, point pattern, sports analytics.

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

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

GÓMEZ RUANO, M. Á., ALARCÓN LÓPEZ, F. and ORTEGA TORO, E. (2015). Analysis of shooting effectiveness in elite basketball according to match status. *Rev. Psicol. Deporte* **24** 0037–41.

GUAN, Y. (2006). A composite likelihood approach in fitting spatial point process models. *J. Amer. Statist. Assoc.* **101** 1502–1512. MR2279475 https://doi.org/10.1198/016214506000000500

GUAN, Y. and SHEN, Y. (2010). A weighted estimating equation approach for inhomogeneous spatial point processes. *Biometrika* **97** 867–880. MR2746157 https://doi.org/10.1093/biomet/asq043

GUDMUNDSSON, J. and HORTON, M. (2017). Spatio-temporal analysis of team sports. *ACM Comput. Surv.* **50** 1–34.

HANNU, O. and JUKKA, N. (1989). Bivariate sign tests. *J. Amer. Statist. Assoc.* **84** 249–259. MR0999686

HU, G., YANG, H.-C. and XUE, Y. (2021). Bayesian group learning for shot selection of professional basketball players. *Stat* **10** Paper No. e324, 12. MR4276018 https://doi.org/10.1002/sta4.324

HU, G., YANG, H.-C., XUE, Y. and DEY, D. K. (2023). Zero-inflated Poisson model with clustered regression coefficients: Application to heterogeneity learning of field goal attempts of professional basketball players. *Canad. J. Statist.* **51** 157–172. MR4551819 https://doi.org/10.1002/cjs.11684

ILLIAN, J., PENTTINEN, A., STOYAN, H. and STOYAN, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. *Statistics in Practice*. Wiley, Chichester. MR2384630

JIAO, J., HU, G. and YAN, J. (2021a). A Bayesian marked spatial point processes model for basketball shot chart. *J. Quant. Anal. Sports* **17** 77–90. https://doi.org/10.1515/jqas-2019-0106

JIAO, J., HU, G. and YAN, J. (2021b). Heterogeneity pursuit for spatial point pattern with application to tree locations: A Bayesian semiparametric recourse. *Environmetrics* **32** Paper No. e2694, 15. MR4323206 https://doi.org/10.1002/env.2694

JUSTEL, A., PEÑA, D. and ZAMAR, R. (1997). A multivariate Kolmogorov–Smirnov test of goodness of fit. *Statist. Probab. Lett.* **35** 251–259. MR1484961 https://doi.org/10.1016/S0167-7152(97)00020-5

KUBATKO, J., OLIVER, D., PELTON, K. and ROSENBAUM, D. T. (2007). A starting point for analyzing basketball statistics. *J. Quant. Anal. Sports* **3** Art. 1, 24. MR2326663 https://doi.org/10.2202/1559-0410.1070

LI, J. and LIU, R. Y. (2004). New nonparametric tests of multivariate locations and scales using data depth. *Statist. Sci.* **19** 686–696. MR2185590 https://doi.org/10.1214/088342304000000594

LIU, R. Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405–414. MR1041400 https://doi.org/10.1214/aos/1176347507

LIU, R. Y., PARELIUS, J. M. and SINGH, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Ann. Statist.* **27** 783–858. With discussion and a rejoinder by Liu and Singh. MR1724033 https://doi.org/10.1214/aos/1018031260

LIU, R. Y. and SINGH, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.* **88** 252–260. MR1212489

LIU, S. and WU, W. (2017). Generalized Mahalanobis depth in point process and its application in neural coding. *Ann. Appl. Stat.* **11** 992–1010. MR3693555 https://doi.org/10.1214/17-AOAS1030

MILLER, A., BORNN, L., ADAMS, R. and GOLDSBERRY, K. (2014). Factorized point process intensities: A spatial analysis of professional basketball. In *International Conference on Machine Learning* 235–243.

MOHD RAZALI, N. and YAP, B. (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *J. Stat. Model. Analytics* **2**.

OJA, H. (1983). Descriptive statistics for multivariate distributions. *Statist. Probab. Lett.* **1** 327–332. MR0721446 https://doi.org/10.1016/0167-7152(83)90054-8

PEACOCK, J. A. (1983). Two-dimensional goodness-of-fit testing in astronomy. *Mon. Not. R. Astron. Soc.* **202** 615–627.

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. and VETTERLING, W. T. (1986). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge Univ. Press, Cambridge. MR0833288

QI, K., CHEN, Y. and WU, W. (2021). Dirichlet depths for point process. *Electron. J. Stat.* **15** 3574–3610. MR4298977 https://doi.org/10.1214/21-ejs1867

QI, K., HU, G. and WU, W. (2024). Supplement to "Are made and missed different? An analysis of field goal attempts of professional basketball players via depth based testing procedure." https://doi.org/10.1214/24-AOAS1899SUPP

REICH, B. J., HODGES, J. S., CARLIN, B. P. and REICH, A. M. (2006). A spatial analysis of basketball shot chart data. *Amer. Statist.* **60** 3–12. MR2224131 https://doi.org/10.1198/000313006X90305

ROUSSEEUW, P. J. and RUTS, I. (1996). Bivariate location depth. *J. R. Stat. Soc., Ser. C* **45** 516–526.

SHORTRIDGE, A., GOLDSBERRY, K. and ADAMS, M. (2014). Creating space to shoot: Quantifying spatial relative field goal efficiency in basketball. *J. Quant. Anal. Sports* **10** 303–313.

TAYLOR, B. M., DAVIES, T. M., ROWLINGSON, B. S. and DIGGLE, P. J. (2013). lgcp: An R package for inference with spatial and spatio-temporal log-Gaussian Cox processes. *J. Stat. Softw.* **52** 1–40.

TUKEY, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians* (*Vancouver, B.C.*, 1974), *Vol.* 2 523–531. Canad. Math. Congr., Montreal, QC. MR0426989

VARDI, Y. and ZHANG, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA* **97** 1423–1426. MR1740461 https://doi.org/10.1073/pnas.97.4.1423

YIN, F., HU, G. and SHEN, W. (2023). Analysis of professional basketball field goal attempts via a Bayesian matrix clustering approach. *J. Comput. Graph. Statist.* **32** 49–60. MR4552936 https://doi.org/10.1080/10618600.2022.2085727

YIN, F., JIAO, J., YAN, J. and HU, G. (2022). Bayesian nonparametric learning for point processes with spatial homogeneity: A spatial analysis of NBA shot locations. In *International Conference on Machine Learning* 25523–25551. PMLR, Baltimore.

ZHANG, C., XIANG, Y. and SHEN, X. (2012). Some multivariate goodness-of-fit tests based on data depth. *J. Appl. Stat.* **39** 385–397. MR2879827 https://doi.org/10.1080/02664763.2011.594033

ZUO, Y. (2003). Projection-based depth functions and associated medians. *Ann. Statist.* **31** 1460–1490. MR2012822 https://doi.org/10.1214/aos/1065705115

ZUO, Y. and HE, X. (2006). On the limiting distributions of multivariate depth-based rank sum statistics and related tests. *Ann. Statist.* **34** 2879–2896. MR2329471 https://doi.org/10.1214/009053606000000876

ZUO, Y. and SERFLING, R. (2000). General notions of statistical depth function. *Ann. Statist.* **28** 461–482. MR1790005 https://doi.org/10.1214/aos/1016218226

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2635–2655
https://doi.org/10.1214/24-AOAS1900

PREDICTION INTERVALS FOR ECONOMIC FIXED-EVENT FORECASTS

BY FABIAN KRÜGER[1,a] AND HENDRIK PLETT[2,b]

[1]*Department of Economics and Management, Karlsruhe Institute of Technology,* [a]*fabian.krueger@kit.edu*
[2]*Master's student at ETH Zürich,* [b]*hendrik.plett@web.de*

The fixed-event forecasting setup is common in economic policy. It involves a sequence of forecasts of the same ("fixed") predictand so that the difficulty of the forecasting problem decreases over time. Fixed-event point forecasts are typically published without a quantitative measure of uncertainty. To construct such a measure, we consider forecast postprocessing techniques tailored to the fixed-event case. We develop regression methods that impose constraints motivated by the problem at hand and use these methods to construct prediction intervals for gross domestic product (GDP) growth in Germany and the U.S.

REFERENCES

ANATOLYEV, S. and GOSPODINOV, N. (2010). Modeling financial return dynamics via decomposition. *J. Bus. Econom. Statist.* **28** 232–245. MR2681198 https://doi.org/10.1198/jbes.2010.07017

ARUOBA, S. B., DIEBOLD, F. X. and SCOTTI, C. (2009). Real-time measurement of business conditions. *J. Bus. Econom. Statist.* **27** 417–427. MR2572030 https://doi.org/10.1198/jbes.2009.07205

BERGMEIR, C., HYNDMAN, R. J. and KOO, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput. Statist. Data Anal.* **120** 70–83. MR3742209 https://doi.org/10.1016/j.csda.2017.11.003

BRACHER, J., RAY, E. L., GNEITING, T. and REICH, N. G. (2021a). Evaluating epidemic forecasts in an interval format. *PLoS Comput. Biol.* **17** e1008618.

BRACHER, J., WOLFFRAM, D., DEUSCHEL, J., GÖRGEN, K., KETTERER, J. L., ULLRICH, A., ABBOTT, S., BARBAROSSA, M. V., BERTSIMAS, D. et al. (2021b). A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. *Nat. Commun.* **12** 5173.

BRAVE, S. A., BUTTERS, R. A. and KELLEY, D. (2019). A new 'big data' index of US economic activity. *Economic Perspectives, Federal Reserve Bank of Chicago* **1**.

CHRISTOFFERSEN, P. F. and DIEBOLD, F. X. (2006). Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Manage. Sci.* **52** 1273–1287.

CLAESKENS, G., MAGNUS, J. R., VASNEV, A. L. and WANG, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *Int. J. Forecast.* **32** 754–762.

CLARK, T. E., GANICS, G. and MERTENS, E. (2022). What is the predictive value of SPF point and density forecasts? Working Paper No. 22–37, Federal Reserve Bank of Cleveland.

CLARK, T. E., MCCRACKEN, M. W. and MERTENS, E. (2020). Modeling time-varying uncertainty of multiple-horizon forecast errors. *Rev. Econ. Stat.* **102** 17–33.

CLEMENTS, M. P. (2010). Explanations of the inconsistencies in survey respondents' forecasts. *Eur. Econ. Rev.* **54** 536–549.

CLEMENTS, M. P. (2014). Forecast uncertainty—*ex ante* and *ex post*: U.S. inflation and output growth. *J. Bus. Econom. Statist.* **32** 206–216. MR3207834 https://doi.org/10.1080/07350015.2013.859618

CLEMENTS, M. P. (2018). Are macroeconomic density forecasts informative? *Int. J. Forecast.* **34** 181–198.

CROUSHORE, D. and STARK, T. (2001). A real-time data set for macroeconomists. *J. Econometrics* **105** 111–130. MR1864214 https://doi.org/10.1016/S0304-4076(01)00072-0

CROUSHORE, D. and STARK, T. (2019). Fifty years of the survey of professional forecasters. *Economic Insights* **4**. Federal Reserve Bank of Philadelphia.

DHAMI, M. K. and MANDEL, D. R. (2022). Communicating uncertainty using words and numbers. *Trends Cogn. Sci.* **26** 514–526. https://doi.org/10.1016/j.tics.2022.03.002

DIEBOLD, F. X. and GÖBEL, M. (2022). A benchmark model for fixed-target Arctic sea ice forecasting. *Econom. Lett.* **215** 110478.

Key words and phrases. Economic forecasting, prediction intervals.

DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **13** 253–263.

DÖPKE, J. and FRITSCHE, U. (2006). Growth and inflation forecasts for Germany: A panel-based assessment of accuracy and efficiency. *Empir. Econ.* **31** 777–798.

ENGELBERG, J., MANSKI, C. F. and WILLIAMS, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *J. Bus. Econom. Statist.* **27** 30–41. MR2484982 https://doi.org/10.1198/jbes.2009.0003

ERASLAN, S. and GÖTZ, T. (2021). An unconventional weekly economic activity index for Germany. *Econom. Lett.* **204** 109881.

EUROPEAN COMMISSION (2022). European economic forecast: Summer 2022. Institutional Paper 183, July 2022.

FAUST, J. and WRIGHT, J. H. (2013). Forecasting inflation. In *Handbook of Economic Forecasting* **2** 2–56. Elsevier, Amsterdam.

FEDERAL RESERVE BANK OF PHILADELPHIA (2021). Survey of professional forecasters: Documentation. Available at https://www.philadelphiafed.org/-/media/frbp/assets/surveys-and-data/survey-of-professional-forecasters/spf-documentation.pdf?la=en&hash=F2D73A2CE0C3EA90E71A363719588D25 (last accessed: September 21, 2022).

FOLTAS, A. and PIERDZIOCH, C. (2022). On the efficiency of German growth forecasts: An empirical analysis using quantile random forests and density forecasts. *Appl. Econ. Lett.* **29** 1644–1653.

GALBRAITH, J. W. and VAN NORDEN, S. (2012). Assessing gross domestic product and inflation probability forecasts derived from Bank of England fan charts. *J. Roy. Statist. Soc. Ser. A* **175** 713–727. MR2948371 https://doi.org/10.1111/j.1467-985X.2011.01012.x

GANICS, G., ROSSI, B. and SEKHPOSYAN, T. (2023). From fixed-event to fixed-horizon density forecasts: Obtaining measures of multi-horizon uncertainty from survey density forecasts. *J. Money Credit Bank.*. To appear.

GENRE, V., KENNY, G., MEYLER, A. and TIMMERMANN, A. (2013). Combining expert forecasts: Can anything beat the simple average? *Int. J. Forecast.* **29** 108–121.

GNEITING, T. (2011). Quantiles as optimal point forecasts. *Int. J. Forecast.* **27** 197–207.

GNEITING, T. and RAFTERY, A. E. (2005). Weather forecasting with ensemble methods. *Science* **310** 248–249. https://doi.org/10.1126/science.1115255

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

GNEITING, T., RAFTERY, A. E., WESTVELD, A. H. and GOLDMAN, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133** 1098–1118.

GNEITING, T. and RANJAN, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econom. Statist.* **29** 411–422. MR2848512 https://doi.org/10.1198/jbes.2010.08110

GNEITING, T. and RANJAN, R. (2013). Combining predictive distributions. *Electron. J. Stat.* **7** 1747–1782. MR3080409 https://doi.org/10.1214/13-EJS823

HALLE INSTITUTE FOR ECONOMIC RESEARCH (2022). IWH Forecasting Dashboard. Available at https://www.iwh-halle.de/ForDas (last accessed: August 14, 2023).

HANSEN, B. E. and LEE, S. (2019). Asymptotic theory for clustered samples. *J. Econometrics* **210** 268–290. MR3958406 https://doi.org/10.1016/j.jeconom.2019.02.001

HEINISCH, K., BEHRENS, C., DÖPKE, J., FOLTAS, A., FRITSCHE, U., KÖHLER, T., MÜLLER, K., PUCKELWALD, J. and REICHMAYR, H. (2023). The IWH forecasting dashboard: From forecasts to evaluation and comparison. *Jahrbücher Für Nationalökonomie und Statistik*. To appear.

HENZI, A. (2023). Consistent estimation of distribution functions under increasing concave and convex stochastic ordering. *J. Bus. Econom. Statist.* **41** 1203–1214. MR4650455 https://doi.org/10.1080/07350015.2022.2116026

HENZI, A., MÖSCHING, A. and DÜMBGEN, L. (2022). Accelerating the pool-adjacent-violators algorithm for isotonic distributional regression. *Methodol. Comput. Appl. Probab.* **24** 2633–2645. MR4528395 https://doi.org/10.1007/s11009-022-09937-2

HENZI, A., ZIEGEL, J. F. and GNEITING, T. (2021). Isotonic distributional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 963–993. MR4349124

JORDAN, A., KRÜGER, F. and LERCH, S. (2019). Evaluating probabilistic forecasts with scoringRules. *J. Stat. Softw.* **90** 1–37.

KNÜPPEL, M. (2014). Efficient estimation of forecast uncertainty based on recent forecast errors. *Int. J. Forecast.* **30** 257–267.

KNÜPPEL, M. and VLADU, A. L. (2016). Approximating fixed-horizon forecasts using fixed-event forecasts. Bundesbank Discussion Paper No. 28/2016.

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

KÖHLER, T. and DÖPKE, J. (2023). Will the last be the first? Ranking German macroeconomic forecasters based on different criteria. *Empir. Econ.* **64** 797–832.

KRÜGER, F., LERCH, S., THORARINSDOTTIR, T. and GNEITING, T. (2021). Predictive inference based on Markov chain Monte Carlo output. *Int. Stat. Rev.* **89** 274–301. MR4411906 https://doi.org/10.1111/insr.12405

KRÜGER, F. and NOLTE, I. (2016). Disagreement versus uncertainty: Evidence from distribution forecasts. *J. Bank. Financ.* **72** S172–S186.

KRÜGER, F. and PAVLOVA, L. (2024). Quantifying subjective uncertainty in survey expectations. *Int. J. Forecast.* **40** 796–810.

KRÜGER, F. and PLETT, H. (2024). Supplement to "Prediction intervals for economic fixed-event forecasts." https://doi.org/10.1214/24-AOAS1900SUPPA, https://doi.org/10.1214/24-AOAS1900SUPPB

KRÜGER, F. and ZIEGEL, J. F. (2021). Generic conditions for forecast dominance. *J. Bus. Econom. Statist.* **39** 972–983. MR4319685 https://doi.org/10.1080/07350015.2020.1741376

LAZARUS, E., LEWIS, D. J., STOCK, J. H. and WATSON, M. W. (2018). HAR inference: Recommendations for practice. *J. Bus. Econom. Statist.* **36** 541–559. MR3871697 https://doi.org/10.1080/07350015.2018.1506926

LEWIS, D. J., MERTENS, K., STOCK, J. H. and TRIVEDI, M. (2022). Measuring real activity using a weekly economic index. *J. Appl. Econometrics* **37** 667–687.

LICHTENDAHL JR, K. C., GRUSHKA-COCKAYNE, Y. and WINKLER, R. L. (2013). Is it better to average probabilities or quantiles? *Manage. Sci.* **59** 1594–1611.

MATHESON, J. E. and WINKLER, R. L. (1976). Scoring rules for continuous probability distributions. *Manage. Sci.* **22** 1087–1096.

PATTON, A. J. and TIMMERMANN, A. (2011). Predictability of output growth and inflation: A multi-horizon survey approach. *J. Bus. Econom. Statist.* **29** 397–410. MR2848511 https://doi.org/10.1198/jbes.2010.08347

PATTON, A. J. and TIMMERMANN, A. (2012). Forecast rationality tests based on multi-horizon bounds. *J. Bus. Econom. Statist.* **30** 1–17. MR2899176 https://doi.org/10.1080/07350015.2012.634337

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

RAFTERY, A. E. (2016). Use and communication of probabilistic forecasts. *Stat. Anal. Data Min.* **9** 397–410. MR3580431 https://doi.org/10.1002/sam.11302

RASP, S. and LERCH, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Mon. Weather Rev.* **146** 3885–3900.

REIFSCHNEIDER, D. and TULIP, P. (2019). Gauging the uncertainty of the economic outlook using historical forecasting errors: The Federal Reserve's approach. *Int. J. Forecast.* **35** 1564–1582.

SHAKED, M. and SHANTHIKUMAR, J. G. (2007). *Stochastic Orders*. *Springer Series in Statistics*. Springer, New York. MR2265633 https://doi.org/10.1007/978-0-387-34675-5

TAGESSCHAU (2022). Konjunkturprognosen für Deutschland. Available at https://www.tagesschau.de/wirtschaft/konjunktur/konjunkturprognose114.html (last accessed: October 6, 2022).

VANNITSEM, S., BREMNES, J. B., DEMAEYER, J., EVANS, G. R., FLOWERDEW, J., HEMRI, S., LERCH, S., ROBERTS, N., THEIS, S. et al. (2021). Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bull. Am. Meteorol. Soc.* **102** E681–E699.

WANG, X., HYNDMAN, R. J., LI, F. and KANG, Y. (2023). Forecast combinations: An over 50-year review. *Int. J. Forecast.* **39** 1518–1547.

ZEILEIS, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *J. Stat. Softw.* **11** 1–17.

ZEILEIS, A., KÖLL, S. and GRAHAM, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *J. Stat. Softw.* **95** 1–36.

The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2656–2677
https://doi.org/10.1214/24-AOAS1901

INTEGRATING MENDELIAN RANDOMIZATION WITH CAUSAL MEDIATION ANALYSES FOR CHARACTERIZING DIRECT AND INDIRECT EXPOSURE-TO-OUTCOME EFFECTS

BY FAN YANG[1,a], LIN S. CHEN[2,b], SHAHRAM OVEISGHARAN[3,c], DAWOOD DARBAR[4,e] AND DAVID A. BENNETT[3,d]

[1] *Yau Mathematical Sciences Center, Tsinghua University,* [a]*yangfan1987@tsinghua.edu.cn*
[2] *Department of Public Health Sciences, The University of Chicago,* [b]*lchen4@bsd.uchicago.edu*
[3] *Rush Alzheimer's Disease Center, Rush University Medical Center,* [c]*Shahram_Oveisgharan@rush.edu,* [d]*David_A_Bennett@rush.edu*
[4] *Department of Pharmacology, University of Illinois at Chicago,* [e]*darbar@uic.edu*

Mendelian randomization (MR) assesses the total effect of exposure on outcome. With the rapidly increasing availability of summary statistics from genome-wide association studies (GWASs), MR leverages existing summary statistics and is widely used to study the causal effects among complex traits and diseases. The total effect in the population is a sum of indirect and direct effects. For complex disease outcomes with complicated etiologies and/or for modifiable exposure traits, there may exist more than one pathway between exposure and outcome. The direct effect and the indirect effect via a mediator of interest could be in opposite directions, and the total effect estimates may not be informative for treatment and prevention decision-making or may even be misleading for different subgroups of patients. Causal mediation analysis delineates the indirect effect of exposure on outcome operating through the mediator and the direct effect transmitted through other mechanisms. However, causal mediation analysis often requires individual-level data measured on exposure, outcome, mediator and confounding variables, and the power of the mediation analysis is restricted by sample size. In this work, motivated by a study of the effects of atrial fibrillation (AF) on Alzheimer's dementia, we propose a framework for Integrative Mendelian randomization and Mediation Analysis (IMMA). The proposed method integrates the total effect estimates from MR analyses based on large-scale GWASs with the direct and indirect effect estimates from mediation analysis based on individual-level data of a limited sample size. We introduce a series of IMMA models under the scenarios with or without exposure-mediator interaction and/or study heterogeneity. The proposed IMMA models improve the estimation and the power of inference on the direct and indirect effects in the population. Our analyses showed a significant positive direct effect of AF on Alzheimer's dementia risk not through the use of the oral anticoagulant treatment and a significant indirect effect of AF-induced anticoagulant treatment in reducing Alzheimer's dementia risk. The results suggested potential Alzheimer's dementia risk prediction and prevention strategies for AF patients and paved the way for future reevaluation of anticoagulant treatment guidelines for AF patients. A sensitivity analysis was conducted to assess the sensitivity of the conclusions to a key assumption of the IMMA approach.

REFERENCES

BENNETT, D. A., BUCHMAN, A. S., BOYLE, P. A., BARNES, L. L., WILSON, R. S. and SCHNEIDER, J. A. (2018). Religious orders study and rush memory and aging project. *J. Alzheimer's Dis.* **64** S161–S189. https://doi.org/10.3233/JAD-179939

BORENSTEIN, M., HEDGES, L. V., HIGGINS, J. P. and ROTHSTEIN, H. R. (2021). *Introduction to Meta-Analysis*. Wiley, New York.

Key words and phrases. Integrative analysis, Mendelian randomization, causal mediation analysis.

BOWDEN, J., SMITH, G. D. and BURGESS, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44** 512–525.

BUNCH, T. J., WEISS, J. P., CRANDALL, B. G., MAY, H. T., BAIR, T. L., OSBORN, J. S., ANDERSON, J. L., MUHLESTEIN, J. B., HORNE, B. D. et al. (2010). Atrial fibrillation is independently associated with senile, vascular, and Alzheimer's dementia. *Heart Rhythm* **7** 433–437.

BUNIELLO, A., MACARTHUR, J. A. L., CEREZO, M., HARRIS, L. W., HAYHURST, J., MALANGONE, C., MCMAHON, A., MORALES, J., MOUNTJOY, E. et al. (2019). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47** D1005–D1012. https://doi.org/10.1093/nar/gky1120

BURGESS, S., BUTTERWORTH, A. and THOMPSON, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37** 658–665. https://doi.org/10.1002/gepi.21758

BURGESS, S., SCOTT, R. A., TIMPSON, N. J., SMITH, G. D. and THOMPSON, S. G. (2015). Using published data in Mendelian randomization: A blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* **30** 543–552.

BURGESS, S., SMALL, D. S. and THOMPSON, S. G. (2017). A review of instrumental variable estimators for Mendelian randomization. *Stat. Methods Med. Res.* **26** 2333–2355. MR3712236 https://doi.org/10.1177/0962280215597579

CRUZ, D., PINTO, R., FREITAS-SILVA, M., NUNES, J. P. and MEDEIROS, R. (2019). GWAS contribution to atrial fibrillation and atrial fibrillation-related stroke: Pathophysiological implications. *Pharmacogenomics* **20** 765–780. https://doi.org/10.2217/pgs-2019-0054

DANIEL, R. M., DE STAVOLA, B. L., COUSENS, S. N. and VANSTEELANDT, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics* **71** 1–14. MR3335344 https://doi.org/10.1111/biom.12248

DIDELEZ, V., MENG, S. and SHEEHAN, N. A. (2010). Assumptions of IV methods for observational epidemiology. *Statist. Sci.* **25** 22–40. MR2741813 https://doi.org/10.1214/09-STS316

FRIBERG, L. and ROSENQVIST, M. (2018). Less dementia with oral anticoagulation in atrial fibrillation. *Eur. Heart J.* **39** 453–460. https://doi.org/10.1093/eurheartj/ehx579

FULCHER, I. R., SHI, X. and TCHETGEN, E. J. T. (2019). Estimation of natural indirect effects robust to unmeasured confounding and mediator measurement error. *Epidemiology* **30** 825–834.

HARING, B., LENG, X., ROBINSON, J., JOHNSON, K. C., JACKSON, R. D., BEYTH, R., WACTAWSKI-WENDE, J., VON BALLMOOS, M. W., GOVEAS, J. S. et al. (2013). Cardiovascular disease and cognitive decline in postmenopausal women: Results from the women's health initiative memory study. *J. Amer. Heart Assoc.* **2**.

HONG, G., YANG, F. and QIN, X. (2023). Posttreatment confounding in causal mediation studies: A cutting-edge problem and a novel solution via sensitivity analysis. *Biometrics* **79** 1042–1056. MR4606335

IHARA, M. and WASHIDA, K. (2018). Linking atrial fibrillation with Alzheimer's disease: Epidemiological, pathological, and mechanistic evidence. *J. Alzheimer's Dis.* **62** 61–72. https://doi.org/10.3233/JAD-170970

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

IMAI, K. and YAMAMOTO, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Polit. Anal.* **21** 141–171.

KALANTARIAN, S., STERN, T. A., MANSOUR, M. and RUSKIN, J. N. (2013). Cognitive impairment associated with atrial fibrillation: A meta-analysis. *Ann. Intern. Med.* **158** 338–346. https://doi.org/10.7326/0003-4819-158-5-201303050-00007

KAWABATA-YOSHIHARA, L. A., SCAZUFCA, M., SANTOS, I. D. S., WHITAKER, A., KAWABATA, V. S., BENSEÑOR, I. M., MENEZES, P. R. and LOTUFO, P. A. (2012). Atrial fibrillation and dementia: Results from the Sao Paulo ageing & health study. *Arq. Bras. Cardiol.* **99** 1108–1114.

KIM, D., YANG, P.-S., SUNG, J.-H., JANG, E., YU, H. T., KIM, T.-H., UHM, J.-S., KIM, J.-Y., PAK, H.-N. et al. (2020). Less dementia after catheter ablation for atrial fibrillation: A nationwide cohort study. *Eur. Heart J.* **41** 4483–4493.

LAMBERT, J. C., IBRAHIM-VERBAAS, C. A., HAROLD, D., NAJ, A. C., SIMS, R., BELLENGUEZ, C., DESTAFANO, A. L., BIS, J. C., BEECHAM, G. W. et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45** 1452–8.

LAWLOR, D. A., HARBORD, R. M., STERNE, J. A. C., TIMPSON, N. and SMITH, G. D. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27** 1133–1163. MR2420151 https://doi.org/10.1002/sim.3034

MIYASAKA, Y., BARNES, M. E., GERSH, B. J., CHA, S. S., BAILEY, K. R., ABHAYARATNA, W. P., SEWARD, J. B. and TSANG, T. S. M. (2006). Secular trends in incidence of atrial fibrillation in Olmsted County, Minnesota, 1980 to 2000, and implications on the projections for future prevalence. *Circulation* **114** 119–125. https://doi.org/10.1161/CIRCULATIONAHA.105.595140

PAN, Y., WANG, Y. and WANG, Y. (2020). Investigation of causal effect of atrial fibrillation on Alzheimer disease: A Mendelian randomization study. *J. Amer. Heart Assoc.* **9**.

PAULE, R. C. and MANDEL, J. (1982). Consensus values and weighting factors. *J. Res. Natl. Bur. Stand.* **87** 377–385. https://doi.org/10.6028/jres.087.022

PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. UAI'*01 411–420. Morgan Kaufmann, San Francisco, CA, USA.

PETERS, R., POULTER, R., BECKETT, N., FORETTE, F., FAGARD, R., POTTER, J., SWIFT, C., ANDERSON, C., FLETCHER, A. et al. (2009). Cardiovascular and biochemical risk factors for incident dementia in the hypertension in the very elderly trial. *J. Hypertens.* **27** 2055–2062. https://doi.org/10.1097/HJH.0b013e32832f4f02

RASTAS, S., VERKKONIEMI, A., POLVIKOSKI, T., JUVA, K., NIINISTÖ, L., MATTILA, K., LÄNSIMIES, E., PIRTTILÄ, T. and SULKAVA, R. (2007). Atrial fibrillation, stroke, and cognition: A longitudinal population-based study of people aged 85 and older. *Stroke* **38** 1454–1460. https://doi.org/10.1161/STROKEAHA.106.477299

REES, J. M. B., WOOD, A. M., DUDBRIDGE, F. and BURGESS, S. (2019). Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates. *PLoS ONE* **14** e0222362. https://doi.org/10.1371/journal.pone.0222362

ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155. https://doi.org/10.1097/00001648-199203000-00013

ROSELLI, C., CHAFFIN, M. D., WENG, L.-C., AESCHBACHER, S., AHLBERG, G., ALBERT, C. M., ALMGREN, P., ALONSO, A., ANDERSON, C. D. et al. (2018). Multi-ethnic genome-wide association study for atrial fibrillation. *Nat. Genet.* **50** 1225–1233.

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

SMITH, G. D. and EBRAHIM, S. (2003). 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32** 1–22. https://doi.org/10.1093/ije/dyg070

SOBEL, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol. Method.* **13** 290–312.

SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. MR1092986

THACKER, E. L., MCKNIGHT, B., PSATY, B. M., LONGSTRETH, W. T. JR., SITLANI, C. M., DUBLIN, S., ARNOLD, A. M., FITZPATRICK, A. L., GOTTESMAN, R. F. et al. (2013). Atrial fibrillation and cognitive decline: A longitudinal cohort study. *Neurology* **81** 119–125. https://doi.org/10.1212/WNL.0b013e31829a33d1

VALERI, L. and VANDERWEELE, T. J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychol. Methods* **18** 137–150. https://doi.org/10.1037/a0031034

VANDERWEELE, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford Univ. Press, London.

XUE, H., SHEN, X. and PAN, W. (2021). Constrained maximum likelihood-based Mendelian randomization robust to both correlated and uncorrelated pleiotropic effects. *Am. J. Hum. Genet.* **108** 1251–1269.

YANG, F., CHEN, L. S., OVEISGHARAN, S., DARBAR, D. and BENNETT, D. A. (2024). Supplement to "Integrating Mendelian randomization with causal mediation analyses for characterizing direct and indirect exposure-to-outcome effects." https://doi.org/10.1214/24-AOAS1901SUPPA, https://doi.org/10.1214/24-AOAS1901SUPPB

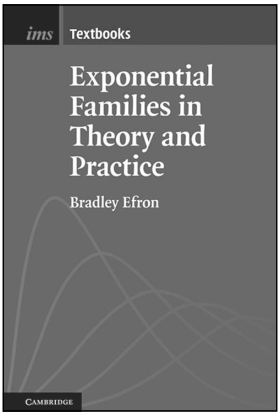
YANG, S. and DING, P. (2020). Combining multiple observational data sources to estimate causal effects. *J. Amer. Statist. Assoc.* **115** 1540–1554. MR4143484 https://doi.org/10.1080/01621459.2019.1609973

ZHAO, Q., WANG, J., HEMANI, G., BOWDEN, J. and SMALL, D. S. (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Ann. Statist.* **48** 1742–1769. MR4124342 https://doi.org/10.1214/19-AOS1866

NEW BOOK

The Institute of Mathematical Statistics presents

IMS TEXTBOOKS



ims Textbooks

Exponential Families in Theory and Practice

Bradley Efron

CAMBRIDGE

Exponential Families in Theory and Practice

Bradley Efron, Stanford University

During the past half-century, exponential families have attained a position at the center of parametric statistical inference. Theoretical advances have been matched, and more than matched, in the world of applications, where logistic regression by itself has become the go-to methodology in medical statistics, computer-based prediction algorithms, and the social sciences. This book is based on a one-semester graduate course for first year Ph.D. and advanced master's students. After presenting the basic structure of univariate and multivariate exponential families, their application to generalized linear models including logistic and Poisson regression is described in detail, emphasizing geometrical ideas, computational practice, and the analogy with ordinary linear regression. Connections are made with a variety of current statistical methodologies: missing data, survival analysis and proportional hazards, false discovery rates, bootstrapping, and empirical Bayes analysis. The book connects exponential family theory with its applications in a way that doesn't require advanced mathematical preparation.

Hardback $105.00

Paperback $39.99

IMS members are entitled to a 40% discount: email ims@imstat.org to request your code

www.imstat.org/cup/

Cambridge University Press, with the Institute of Mathematical Statistics, established the *IMS Monographs* and *IMS Textbooks* series of high-quality books. The series editors are Mark Handcock, Ramon van Handel, Arnaud Doucet, and John Aston.