

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

Function-on-scalar quantile regression with application to mass spectrometry proteomics data	YUSHA LIU, MENG LI AND JEFFREY S. MORRIS	521
Compression of climate simulations with a nonstationary global SpatioTemporal SPDE model	GEIR-ARNE FUGLSTAD AND STEFANO CASTRUCCIO	542
An efficient and computationally robust statistical method for analyzing case-control mother-offspring pair genetic association studies	HONG ZHANG, BHRAMAR MUKHERJEE, VICTORIA ARTHUR, GANG HU, HAGIT HOCHNER AND JINBO CHEN	560
Baseline drift estimation for air quality data using quantile trend filtering	HALLEY L. BRANTLEY, JOSEPH GUINNESS AND ERIC C. CHI	585
A fast particle-based approach for calibrating a 3-D model of the Antarctic ice sheet	BEN SEIYON LEE, MURALI HARAN, ROBERT W. FULLER, DAVID POLLARD AND KLAUS KELLER	605
Torus graphs for multivariate phase coupling analysis	NATALIE KLEIN, JOSUE ORELLANA, SCOTT L. BRINCAT, EARL K. MILLER AND ROBERT E. KASS	635
The stratified micro-randomized trial design: Sample size considerations for testing nested causal effects of time-varying treatments	WALTER DEMPSEY, PENG LIAO, SANTOSH KUMAR AND SUSAN A. MURPHY	661
Accounting for uncertainty about past values in probabilistic projections of the total fertility rate for most countries	PEIRAN LIU AND ADRIAN E. RAFTERY	685
Estimation of dyadic characteristics of family networks using sample survey data	CHRIS SKINNER AND FIONA STEELE	706
Sequential importance sampling for multiresolution Kingman–Tajima coalescent counting	LORENZO CAPPELLO AND JULIA A. PALACIOS	727
Multiview cluster aggregation and splitting, with an application to multiomic breast cancer data	ANTOINE GODICHON-BAGGIONI, CATHY MAUGIS-RABUSSEAU AND ANDREA RAU	752
Semiparametric Bayesian Markov analysis of personalized benefit–risk assessment	DONGYAN YAN, SUBHARUP GUHA, CHUL AHN AND RAM TIWARI	768
Estimation and inference in metabolomics with nonrandom missing data and latent factors	CHRIS MCKENNAN, CAROLE OBER AND DAN NICOLAE	789
Bayesian variable selection for survival data using inverse moment priors	AMIR NIKOOIENEJAD, WENYI WANG AND VALEN E. JOHNSON	809
Evidence factors in a case-control study with application to the effect of flexible sigmoidoscopy screening on colorectal cancer	BIKRAM KARMAKAR, CHYKE A. DOUBENI AND DYLAN S. SMALL	829
A causal exposure response function with local adjustment for confounding: Estimating health effects of exposure to low levels of ambient fine particulate matter	GEORGIA PAPADOGEORGOU AND FRANCESCA DOMINICI	850

continued

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—*Continued from front cover*

- Focused model selection for linear mixed models with an application to whale ecology
CÉLINE CUNEN, LARS WALLØE AND NILS LID HJORT 872
- A global-local approach for detecting hotspots in multiple-response regression
HÉLÈNE RUFFIEUX, ANTHONY C. DAVISON, JÖRG HAGER, JAMIE INSHAW,
BENJAMIN P. FAIRFAX, SYLVIA RICHARDSON AND LEONARDO BOTTOLO 905
- Seasonal warranty prediction based on recurrent event data
QIANQIAN SHAN, YILI HONG AND WILLIAM Q. MEEKER 929
- Generalized accelerated recurrence time model in the presence of a dependent terminal
event BO WEI, ZHUMIN ZHANG, HUICHUAN J. LAI AND LIMIN PENG 956
- Early identification of an impending rockslide location via a spatially-aided Gaussian
mixture model . . . SHUO ZHOU, HOWARD BONDELL, ANTOINETTE TORDSILLAS,
BENJAMIN I. P. RUBINSTEIN AND JAMES BAILEY 977
- A random effects stochastic block model for joint community detection in multiple
networks with applications to neuroimaging SUBHADEEP PAUL AND
YUGUO CHEN 993
- Analyses of preventive care measures with incomplete historical data in electronic
medical records: An example from colorectal cancer screening . . . YINGYE ZHENG,
DOUGLAS A. CORLEY, CHYKE DOUBENI, ETHAN HALM,
SUSAN M. SHORTREED, WILLIAM E. BARLOW, ANN ZAUBER,
TOR DEVIN TOSTESON AND JESSICA CHUBAK 1030
- Accounting for dependent errors in predictors and time-to-event outcomes using
electronic health records, validation samples and multiple imputation
MARK J. GIGANTI, PAMELA A. SHAW, GUANHUA CHEN, SALLY S. BEBAWY,
MEGAN M. TURNER, TIMOTHY R. STERLING AND BRYAN E. SHEPHERD 1045

THE ANNALS OF APPLIED STATISTICS

Vol. 14, No. 2, pp. 521–1061 June 2020

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

IMS OFFICERS

President: Susan Murphy, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

President-Elect: Regina Y. Liu, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854-8019, USA

Past President: Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

Executive Secretary: Edsel Peña, Department of Statistics, University of South Carolina, Columbia, South Carolina 29208-001, USA

Treasurer: Zhengjun Zhang, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706-1510, USA

Program Secretary: Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027-5927, USA

IMS PUBLICATIONS

The Annals of Statistics. *Editors:* Richard J. Samworth, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge, CB3 0WB, UK. Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027, USA

The Annals of Applied Statistics. *Editor-In-Chief:* Karen Kafadar, Department of Statistics, University of Virginia, Charlottesville, VA 22904-4135, USA

The Annals of Probability. *Editor:* Amir Dembo, Department of Statistics and Department of Mathematics, Stanford University, Stanford, California 94305, USA

The Annals of Applied Probability. *Editors:* François Delarue, Laboratoire J. A. Dieudonné, Université de Nice Sophia-Antipolis, France-06108 Nice Cedex 2. Peter Friz, Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany and Weierstrass-Institut für Angewandte Analysis und Stochastik, 10117 Berlin, Germany

Statistical Science. *Editor:* Sonia Petrone, Department of Decision Sciences, Università Bocconi, 20100 Milano MI, Italy

The IMS Bulletin. *Editor:* Vlada Limic, UMR 7501 de l'Université de Strasbourg et du CNRS, 7 rue René Descartes, 67084 Strasbourg Cedex, France

The Annals of Applied Statistics [ISSN 1932-6157 (print); ISSN 1941-7330 (online)], Volume 14, Number 2, June 2020. Published quarterly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, Ohio 44122, 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, 9650 Rockville Pike, Suite L 2310, Bethesda, Maryland 20814-3998, USA.

FUNCTION-ON-SCALAR QUANTILE REGRESSION WITH APPLICATION TO MASS SPECTROMETRY PROTEOMICS DATA

BY YUSHA LIU^{1,*}, MENG LI^{1,†} AND JEFFREY S. MORRIS²

¹Department of Statistics, Rice University, *yl95@rice.edu; †meng@rice.edu

²Division of Biostatistics, University of Pennsylvania, Jeffrey.Morris@penncmedicine.upenn.edu

Mass spectrometry proteomics, characterized by spiky, spatially heterogeneous functional data, can be used to identify potential cancer biomarkers. Existing mass spectrometry analyses utilize mean regression to detect spectral regions that are differentially expressed across groups. However, given the interpatient heterogeneity that is a key hallmark of cancer, many biomarkers are only present at aberrant levels for a subset of, not all, cancer samples. Differences in these biomarkers can easily be missed by mean regression but might be more easily detected by quantile-based approaches. Thus, we propose a unified Bayesian framework to perform quantile regression on functional responses. Our approach utilizes an asymmetric Laplace working likelihood, represents the functional coefficients with basis representations which enable borrowing of strength from nearby locations and places a global-local shrinkage prior on the basis coefficients to achieve adaptive regularization. Different types of basis transform and continuous shrinkage priors can be used in our framework. A scalable Gibbs sampler is developed to generate posterior samples that can be used to perform Bayesian estimation and inference while accounting for multiple testing. Our framework performs quantile regression and coefficient regularization in a unified manner, allowing them to inform each other and leading to improvement in performance over competing methods, as demonstrated by simulation studies. We also introduce an adjustment procedure to the model to improve its frequentist properties of posterior inference. We apply our model to identify proteomic biomarkers of pancreatic cancer that are differentially expressed for a subset of cancer patients compared to the normal controls which were missed by previous mean-regression based approaches. Supplementary Material for this article is available online.

REFERENCES

- BAGGERLY, K. A., MORRIS, J. S. and COOMBES, K. R. (2004). Reproducibility of SELDI-TOF protein patterns in serum: Comparing datasets from different experiments. *Bioinformatics* **20** 777–785.
- BAGGERLY, K. A., MORRIS, J. S., WANG, J., GOLD, D., XIAO, L.-C. and COOMBES, K. R. (2003). A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics* **3** 1667–1672.
- BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *J. Amer. Statist. Assoc.* **110** 1479–1490. [MR3449048 https://doi.org/10.1080/01621459.2014.960967](https://doi.org/10.1080/01621459.2014.960967)
- BROCKHAUS, S. and RUEGAMER, D. (2017). FDboost: Boosting functional regression models.
- BROCKHAUS, S., SCHEIPL, F., HOTHORN, T. and GREVEN, S. (2015). The functional linear array model. *Stat. Model.* **15** 279–300. [MR3349797 https://doi.org/10.1177/1471082X14566913](https://doi.org/10.1177/1471082X14566913)
- CAI, Z. and XU, X. (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *J. Amer. Statist. Assoc.* **103** 1595–1608. [MR2504207 https://doi.org/10.1198/016214508000000977](https://doi.org/10.1198/016214508000000977)
- CARDOT, H., CRAMBES, C. and SARDA, P. (2005). Quantile regression when the covariates are functions. *J. Nonparametr. Stat.* **17** 841–856. [MR2180369 https://doi.org/10.1080/10485250500303015](https://doi.org/10.1080/10485250500303015)
- CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics* 73–80.

Key words and phrases. Bayesian hierarchical model, functional data analysis, functional response regression, global-local shrinkage, proteomic biomarker, quantile regression.

- 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](https://doi.org/10.1093/biomet/asq017)
- 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](https://doi.org/10.1111/j.1467-9868.2011.01008.x)
- COOMBES, K. R., TSAVACHIDIS, S., MORRIS, J. S., BAGGERLY, K. A., HUNG, M.-C. and KUERER, H. M. (2005). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics* **5** 4107–4117.
- DEUTSCH, E. W., LAM, H. and AEBERSOLD, R. (2008). Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol. Genomics* **33** 18–25. <https://doi.org/10.1152/physiolgenomics.00298.2007>
- FASIOLO, M., GOUDE, Y., NEDELLEC, R. and WOOD, S. N. (2018). Fast calibrated additive quantile regression. Preprint. Available at [arXiv:1707.03307](https://arxiv.org/abs/1707.03307).
- FENG, X. and ZHU, L. (2016). Estimation and testing of varying coefficients in quantile regression. *J. Amer. Statist. Assoc.* **111** 266–274. [MR3494658 https://doi.org/10.1080/01621459.2014.1001068](https://doi.org/10.1080/01621459.2014.1001068)
- FERRATY, F., RABHI, A. and VIEU, P. (2005). Conditional quantiles for dependent functional data with application to the climatic El Niño phenomenon. *Sankhyā* **67** 378–398. [MR2208895 https://doi.org/10.2307/2346895](https://doi.org/10.2307/2346895)
- GASTEIGER, E., HOOGLAND, C., GATTIKER, A., WILKINS, M. R., APPEL, R. D., BAIROCH, A. et al. (2005). Protein identification and analysis tools on the ExPASy server. In *The Proteomics Protocols Handbook* 571–607. Springer, Berlin.
- GERACI, M. and BOTTAI, M. (2006). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8** 140–154.
- GUO, J.-C., LI, J., ZHOU, L., YANG, J.-Y., ZHANG, Z.-G., LIANG, Z.-Y., ZHOU, W.-X., YOU, L., ZHANG, T.-P. et al. (2016). CXCL12–CXCR7 axis contributes to the invasive phenotype of pancreatic cancer. *Oncotarget* **7** 62006–62018.
- INNOCENTI, F., OWZAR, K., COX, N. L., EVANS, P., KUBO, M., ZEMBUTSU, H., JIANG, C., HOLLIS, D., MUSHIRODA, T. et al. (2012). A genome-wide association study of overall survival in pancreatic cancer patients treated with gemcitabine in CALGB 80303. *Clin. Cancer Res.* **18** 577–584.
- JAMES, G. M., WANG, J. and ZHU, J. (2009). Functional linear regression that’s interpretable. *Ann. Statist.* **37** 2083–2108. [MR2543686 https://doi.org/10.1214/08-AOS641](https://doi.org/10.1214/08-AOS641)
- KATO, K. (2012). Estimation in functional linear quantile regression. *Ann. Statist.* **40** 3108–3136. [MR3097971 https://doi.org/10.1214/12-AOS1066](https://doi.org/10.1214/12-AOS1066)
- KIM, M.-O. (2007). Quantile regression with varying coefficients. *Ann. Statist.* **35** 92–108. [MR2332270 https://doi.org/10.1214/009053606000000966](https://doi.org/10.1214/009053606000000966)
- KINTER, M. and SHERMAN, N. E. (2005). *Protein Sequencing and Identification Using Tandem Mass Spectrometry* **9**. Wiley, New York.
- KOENKER, R. (2005). *Quantile Regression. Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. [MR2268657 https://doi.org/10.1017/CBO9780511754098](https://doi.org/10.1017/CBO9780511754098)
- KOENKER, R. (2017). quantreg: Quantile regression. R package version 5.33.
- KOENKER, R. and BASSETT, G. JR. (1978). Regression quantiles. *Econometrica* **46** 33–50. [MR0474644 https://doi.org/10.2307/1913643](https://doi.org/10.2307/1913643)
- KOOMEN, J. M., SHIH, L. N., COOMBES, K. R., LI, D., XIAO, L., FIDLER, I. J., ABBRUZZESE, J. L. and KOBAYASHI, R. (2005). Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. *Clin. Cancer Res.* **11** 1110–1118.
- LI, M., WANG, K., MAITY, A. and STAICU, A.-M. (2016). Inference in functional linear quantile regression. Preprint. Available at [arXiv:1602.08793](https://arxiv.org/abs/1602.08793).
- LIAO, H., MOSCHIDIS, E., RIBA-GARCIA, I., ZHANG, Y., UNWIN, R. D., MORRIS, J. S., GRAHAM, J. and DOWSEY, A. W. (2014). A new paradigm for clinical biomarker discovery and screening with mass spectrometry through biomedical image analysis principles. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)* 1332–1335. IEEE, Piscataway, NJ.
- LIU, Y., LI, M. and MORRIS, J. S. (2020). Supplement to “Function-on-scalar quantile regression with application to mass spectrometry proteomics data.” <https://doi.org/10.1214/19-AOAS1319SUPPA>, <https://doi.org/10.1214/19-AOAS1319SUPPB>.
- LUM, K. and GELFAND, A. E. (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Anal.* **7** 235–258. [MR2934947 https://doi.org/10.1214/12-BA708](https://doi.org/10.1214/12-BA708)
- MATLAB (2016). Version 9.1 (R2016b). The MathWorks Inc., Natick, MA.
- MEYER, M. J., COULL, B. A., VERSACE, F., CINCIRIPINI, P. and MORRIS, J. S. (2015). Bayesian function-on-function regression for multilevel functional data. *Biometrics* **71** 563–574. [MR3402592 https://doi.org/10.1111/biom.12299](https://doi.org/10.1111/biom.12299)

- MORRIS, J. S. (2012). Statistical methods for proteomic biomarker discovery based on feature extraction or functional modeling approaches. *Stat. Interface* **5** 117–135. [MR2896986](#) <https://doi.org/10.4310/SII.2012.v5.n1.a11>
- MORRIS, J. S. (2015). Functional regression. *Annu. Rev. Stat. Appl.* **2** 321–359.
- MORRIS, J. S. (2017). Comparison and contrast of two general functional regression modelling frameworks [Discussion of MR3619335]. *Stat. Model.* **17** 59–85. [MR3619339](#) <https://doi.org/10.1177/1471082X16681875>
- MORRIS, J. S. and CARROLL, R. J. (2006). Wavelet-based functional mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 179–199. [MR2188981](#) <https://doi.org/10.1111/j.1467-9868.2006.00539.x>
- 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>
- POLSON, N. G. and SCOTT, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics* 9 501–538. Oxford Univ. Press, Oxford. [MR3204017](#) <https://doi.org/10.1093/acprof:oso/9780199694587.003.0017>
- REED, C. and YU, K. (2009). A partially collapsed Gibbs sampler for Bayesian quantile regression.
- RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. *Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. [MR1998720](#) <https://doi.org/10.1017/CBO9780511755453>
- SORACE, J. M. and ZHAN, M. (2003). A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinform.* **4** Art. ID 24. <https://doi.org/10.1186/1471-2105-4-24>
- SRIRAM, K. (2015). A sandwich likelihood correction for Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Statist. Probab. Lett.* **107** 18–26. [MR3412751](#) <https://doi.org/10.1016/j.spl.2015.07.035>
- SYRING, N. and MARTIN, R. (2019). Calibrating general posterior credible regions. *Biometrika* **106** 479–486. [MR3949316](#) <https://doi.org/10.1093/biomet/asy054>
- R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- 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. J., ZHU, Z. and ZHOU, J. (2009). Quantile regression in partially linear varying coefficient models. *Ann. Statist.* **37** 3841–3866. [MR2572445](#) <https://doi.org/10.1214/09-AOS695>
- XI, R., LI, Y. and HU, Y. (2016). Bayesian quantile regression based on the empirical likelihood with spike and slab priors. *Bayesian Anal.* **11** 821–855. [MR3543910](#) <https://doi.org/10.1214/15-BA975>
- YANG, Y. and HE, X. (2012). Bayesian empirical likelihood for quantile regression. *Ann. Statist.* **40** 1102–1131. [MR2985945](#) <https://doi.org/10.1214/12-AOS1005>
- YANG, Y., WANG, H. J. and HE, X. (2016). Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. *Int. Stat. Rev.* **84** 327–344. [MR3580414](#) <https://doi.org/10.1111/insr.12114>
- YEE, N. S., CHAN, A. S., YEE, J. D. and YEE, R. K. (2012). TRPM7 and TRPM8 ion channels in pancreatic adenocarcinoma: Potential roles as cancer biomarkers and targets. *Scientifica* **2012** Art. ID 415158. <https://doi.org/10.6064/2012/415158>
- YU, K. and MOYED, R. A. (2001). Bayesian quantile regression. *Statist. Probab. Lett.* **54** 437–447. [MR1861390](#) [https://doi.org/10.1016/S0167-7152\(01\)00124-9](https://doi.org/10.1016/S0167-7152(01)00124-9)
- YUE, Y. R. and RUE, H. (2011). Bayesian inference for additive mixed quantile regression models. *Comput. Statist. Data Anal.* **55** 84–96. [MR2736538](#) <https://doi.org/10.1016/j.csda.2010.05.006>
- ZHANG, J., GONZALEZ, E., HESTILOW, T., HASKINS, W. and HUANG, Y. (2009). Review of peak detection algorithms in liquid-chromatography–mass spectrometry. *Curr. Genomics* **10** 388–401. <https://doi.org/10.2174/138920209789177638>

COMPRESSION OF CLIMATE SIMULATIONS WITH A NONSTATIONARY GLOBAL SPATIOTEMPORAL SPDE MODEL

BY GEIR-ARNE FUGLSTAD¹ AND STEFANO CASTRUCCIO²

¹*Department of Mathematical Sciences, Norwegian University of Science and Technology, geir-arne.fuglstad@ntnu.no*

²*Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, scastruc@nd.edu*

Modern climate models pose an ever-increasing storage burden to computational facilities, and the upcoming generation of global simulations from the next Intergovernmental Panel on Climate Change will require a substantial share of the budget of research centers worldwide to be allocated just for this task. A statistical model can be used as a means to mitigate the storage burden by providing a stochastic approximation of the climate simulations. Indeed, if a suitably validated statistical model can be formulated to draw realizations whose spatiotemporal structure is similar to that of the original computer simulations, then the estimated parameters are effectively all the information that needs to be stored. In this work we propose a new statistical model defined via a stochastic partial differential equation (SPDE) on the sphere and in evolving time. The model is able to capture nonstationarities across latitudes, longitudes and land/ocean domains for more than 300 million data points while also overcoming the fundamental limitations of current global statistical models available for compression. Once the model is trained, surrogate runs can be instantaneously generated on a laptop by storing just 20 Megabytes of parameters as opposed to more than six Gigabytes of the original ensemble.

REFERENCES

- BAKER, A. H., XU, H., DENNIS, J. M., LEVY, M. N., NYCHKA, D., MICKELSON, S. A., EDWARDS, J., VERTENSTEIN, M. and WEGENER, A. (2014). A methodology for evaluating the impact of data compression on climate simulation data. In *Proceedings of the 23rd International Symposium on High-Performance Parallel and Distributed Computing, ACM HPDC '14* 203–214.
- BAKER, A. H., HAMMERLING, D. M., MICKELSON, S. A., XU, H., STOLPE, M. B., NAVEAU, P., SANDERSON, B., EBERT-UPHOFF, I., SAMARASINGHE, S. et al. (2016). Evaluating lossy data compression on climate simulation data within a large ensemble. *Geosci. Model Dev.* **9** 4381–4403.
- BAKER, A. H., XU, H., HAMMERLING, D. M., LI, S.-M. and CLYNE, J. P. (2017). Toward a multi-method approach: Lossy data compression for climate simulation data. In *High Performance Computing: ISC High Performance 2017* (J. Kunkel, R. Yokota, M. Taufer and J. Shalf, eds.) 30–42. Springer, Berlin.
- BAKKA, H., VANHATALO, J., ILLIAN, J. B., SIMPSON, D. and RUE, H. (2019). Non-stationary Gaussian models with physical barriers. *Spat. Stat.* **29** 268–288. MR3903698 <https://doi.org/10.1016/j.spasta.2019.01.002>
- BOLIN, D. and LINDGREN, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Ann. Appl. Stat.* **5** 523–550. MR2810408 <https://doi.org/10.1214/10-AOAS383>
- CASTRUCCIO, S. and GENTON, M. G. (2016). Compressing an ensemble with statistical models: An algorithm for global 3D spatio-temporal temperature. *Technometrics* **58** 319–328. MR3520661 <https://doi.org/10.1080/00401706.2015.1027068>
- CASTRUCCIO, S. and GUINNESS, J. (2017). An evolutionary spectrum approach to incorporate large-scale geographical descriptors on global processes. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 329–344. MR3611690 <https://doi.org/10.1111/rssc.12167>
- CASTRUCCIO, S. and STEIN, M. L. (2013). Global space-time models for climate ensembles. *Ann. Appl. Stat.* **7** 1593–1611. MR3127960 <https://doi.org/10.1214/13-AOAS656>
- CASTRUCCIO, S., HU, Z., SANDERSON, B., KARSPECK, A. and HAMMERLING, D. (2019). Reproducing internal variability with few ensemble runs. *J. Climate* **32** 8511–8522.

- EDWARDS, M., HAMMERLING, D. and CASTRUCCIO, S. (2020). Marginally parametrized spatio-temporal models and stepwise maximum likelihood estimation. Available at [arXiv:1806.11388](https://arxiv.org/abs/1806.11388).
- EYRING, V., BONY, S., MEEHL, G. A., SENIOR, C. A., STEVENS, B., STOUFFER, R. J. and TAYLOR, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9** 1937–1958.
- FUENTES, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *J. Amer. Statist. Assoc.* **102** 321–331. [MR2345545 https://doi.org/10.1198/016214506000000852](https://doi.org/10.1198/016214506000000852)
- FUGLSTAD, G.-A. and CASTRUCCIO, S. (2020). Supplement to “Compression of climate simulations with a nonstationary global SpatioTemporal SPDE model.” <https://doi.org/10.1214/20-AOAS1340SUPPA>, <https://doi.org/10.1214/20-AOAS1340SUPPB>
- FUGLSTAD, G.-A., LINDGREN, F., SIMPSON, D. and RUE, H. (2015a). Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Statist. Sinica* **25** 115–133. [MR3328806](https://doi.org/10.1007/s00180-015-0531-1)
- FUGLSTAD, G.-A., SIMPSON, D., LINDGREN, F. and RUE, H. (2015b). Does non-stationary spatial data always require non-stationary random fields? *Spat. Stat.* **14** 505–531. [MR3431054 https://doi.org/10.1016/j.spasta.2015.10.001](https://doi.org/10.1016/j.spasta.2015.10.001)
- FUGLSTAD, G.-A., SIMPSON, D., LINDGREN, F. and RUE, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *J. Amer. Statist. Assoc.* **114** 445–452. [MR3941267 https://doi.org/10.1080/01621459.2017.1415907](https://doi.org/10.1080/01621459.2017.1415907)
- GUINNESS, J. and HAMMERLING, D. (2018). Compression and conditional emulation of climate model output. *J. Amer. Statist. Assoc.* **113** 56–67. [MR3803439 https://doi.org/10.1080/01621459.2017.1395339](https://doi.org/10.1080/01621459.2017.1395339)
- INGEBRIGTSEN, R., LINDGREN, F. and STEINSLAND, I. (2014). Spatial models with explanatory variables in the dependence structure. *Spat. Stat.* **8** 20–38. [MR3326819 https://doi.org/10.1016/j.spasta.2013.06.002](https://doi.org/10.1016/j.spasta.2013.06.002)
- INGEBRIGTSEN, R., LINDGREN, F., STEINSLAND, I. and MARTINO, S. (2015). Estimation of a non-stationary model for annual precipitation in southern Norway using replicates of the spatial field. *Spat. Stat.* **14** 338–364. [MR3431045 https://doi.org/10.1016/j.spasta.2015.07.003](https://doi.org/10.1016/j.spasta.2015.07.003)
- JEONG, J., CASTRUCCIO, S., CRIPPA, P. and GENTON, M. G. (2018). Reducing storage of global wind ensembles with stochastic generators. *Ann. Appl. Stat.* **12** 490–509. [MR3773402 https://doi.org/10.1214/17-AOAS1105](https://doi.org/10.1214/17-AOAS1105)
- JEONG, J., YAN, Y., CASTRUCCIO, S. and GENTON, M. G. (2019). A stochastic generator of global monthly wind energy with Tukey *g*-and-*h* autoregressive processes. *Statist. Sinica* **29** 1105–1126. [MR3932511](https://doi.org/10.1007/s00180-019-00000-1)
- JUN, M. and STEIN, M. L. (2007). An approach to producing space-time covariance functions on spheres. *Technometrics* **49** 468–479. [MR2394558 https://doi.org/10.1198/0040170070000000155](https://doi.org/10.1198/0040170070000000155)
- KAY, J. E., DESER, C., PHILLIPS, A., MAI, A., HANNAY, C., STRAND, G., ARBLASTER, J. M., BATES, S. C., DANABASOGLU, G. et al. (2015). The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Am. Meteorol. Soc.* **96** 1333–1349.
- LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. [MR2853727 https://doi.org/10.1111/j.1467-9868.2011.00777.x](https://doi.org/10.1111/j.1467-9868.2011.00777.x)
- MEEHL, G. A., MOSS, R., TAYLOR, K. E., EYRING, V., STOUFFER, R. J., BONY, S. and STEVENS, B. (2014). Climate model intercomparisons: Preparing for the next phase. *Eos Trans. AGU* **95** 77–78.
- NIKIEMA, O. and LAPRISE, R. (2011). Budget study of the internal variability in ensemble simulations of the Canadian Regional Climate Model at the seasonal scale. *J. Geophys. Res., Atmos.* **116**.
- ROZANOV, JU. A. (1977). Markovian random fields, and stochastic partial differential equations. *Math. USSR, Sb.* **32** 515–534.
- RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields: Theory and Applications. Monographs on Statistics and Applied Probability* **104**. CRC Press/CRC, Boca Raton, FL. [MR2130347 https://doi.org/10.1201/9780203492024](https://doi.org/10.1201/9780203492024)
- SAMPSON, P. D. and GUTTORP, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87** 108–119. <https://doi.org/10.1080/01621459.1992.10475181>
- SMALL, R. J., BACMEISTER, J., BAILEY, D., BAKER, A., BISHOP, S., BRYAN, F., CARON, J., DENNIS, J., GENT, P. et al. (2014). A new synoptic scale resolving global climate simulation using the Community Earth System Model. *J. Adv. Model. Earth Syst.* **6** 1065–1094.
- STEIN, M. L. (2005). Statistical methods for regular monitoring data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 667–687. [MR2210686 https://doi.org/10.1111/j.1467-9868.2005.00520.x](https://doi.org/10.1111/j.1467-9868.2005.00520.x)
- SUN, Y. and GENTON, M. G. (2011). Functional boxplots. *J. Comput. Graph. Statist.* **20** 316–334. [MR2847798 https://doi.org/10.1198/jcgs.2011.09224](https://doi.org/10.1198/jcgs.2011.09224)
- TAYLOR, K., STOUFFER, R. and MEEHL, G. (2012). An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93** 485–498.

- TORN, R. D. (2016). Evaluation of atmosphere and ocean initial condition uncertainty and stochastic exchange coefficients on ensemble tropical cyclone intensity forecasts. *Mon. Weather Rev.* **144** 3487–3506.
- VAN VUUREN, D. P., EDMONDS, J., KAINUMA, M., RIAHI, K., THOMSON, A., HIBBARD, K., HURTT, G. C., KRAM, T., KREY, V. et al. (2011). The representative concentration pathways: An overview. *Clim. Change* **109** 5–31.
- WHITTLE, P. (1954). On stationary processes in the plane. *Biometrika* **41** 434–449. [MR0067450
https://doi.org/10.1093/biomet/41.3-4.434](https://doi.org/10.1093/biomet/41.3-4.434)
- YOSHIMORI, M., STOCKER, T. F., RAIBLE, C. C. and RENOLD, M. (2005). Externally forced and internal variability in ensemble climate simulations of the Maunder minimum. *J. Climate* **18** 4253–4270.

AN EFFICIENT AND COMPUTATIONALLY ROBUST STATISTICAL METHOD FOR ANALYZING CASE-CONTROL MOTHER–OFFSPRING PAIR GENETIC ASSOCIATION STUDIES

BY HONG ZHANG¹, BHARAMAR MUKHERJEE², VICTORIA ARTHUR^{3,*}, GANG HU⁴,
HAGIT HOCHNER⁵ AND JINBO CHEN^{3,**}

¹Department of Statistics and Finance, School of Management, University of Science and Technology of China,
zhangh@ustc.edu.cn

²Department of Biostatistics, University of Michigan, bhramar@umich.edu

³Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania,
*vlynn@pennmedicine.upenn.edu; **jinboche@mail.med.upenn.edu

⁴Chronic Disease Epidemiology Laboratory, Pennington Biomedical Research Center, gang.hu@pbrc.edu

⁵The Hebrew University of Jerusalem, hagith@ekmd.huji.ac.il

Case-control mother–offspring pair design has been widely adopted for studying early-life and women’s pregnancy health. It allows assessment of pre- and perinatal environmental risk factors as well as both maternal and offspring genetic risk factors. Data arising from this design is routinely analyzed using standard prospective logistic regression. Such data has two unique features: the offspring genotypes are not correlated with maternal environmental risk factors given maternal genotypes, and offspring and maternal genotypes are related through mendelian transmission. In this work, built upon a novel regression model relating maternal genotypes to environmental risk factors, we proposed a novel retrospective likelihood method that effectively utilized the two data features to increase statistical efficiency for detecting maternal and offspring genetic effects. The inference procedure was based on a profile likelihood derived using the Lagrange multiplier method, but we replaced the multipliers with their large sample limits to enable highly efficient and computationally stable estimation. We showed that our proposed estimates of odds ratio association parameters are consistent and asymptotically normally distributed and demonstrated the finite sample performance through extensive simulation studies and application to genetic association studies of birth weight and gestational diabetes mellitus.

REFERENCES

- ABUBAKARI, A., KYNAST-WOLF, G. and JAHN, A. (2015). Maternal determinants of birth weight in northern Ghana. *PLoS ONE* **10** e0135641. <https://doi.org/10.1371/journal.pone.0135641>
- AGRESTI, A. (2013). *Categorical Data Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. [MR3087436](#)
- AINSWORTH, H. F., UNWIN, J., JAMISON, D. L. and CORDELL, H. J. (2011a). Investigation of maternal effects, maternal-foetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. *Genet. Epidemiol.* **35** 19–45.
- AINSWORTH, H. F., UNWIN, J., JAMISON, D. L. and CORDELL, H. J. (2011b). Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. *Genet. Epidemiol.* **35** 19–45. <https://doi.org/10.1002/gepi.20547>
- CHATTERJEE, N. and CARROLL, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92** 399–418. [MR2201367](#) <https://doi.org/10.1093/biomet/92.2.399>
- CHEN, J., LIN, D. and HOCHNER, H. (2012). Semiparametric maximum likelihood methods for analyzing genetic and environment effects with case-control mother-child pair data. *Biometrics* **68** 869–877. [MR3055191](#) <https://doi.org/10.1111/j.1541-0420.2011.01728.x>

Key words and phrases. Case-control mother–offspring pair design, genetic association, profile likelihood, saddle point problem, retrospective likelihood.

- CHEN, J., ZHENG, H. and WILSON, M. L. (2009). Likelihood ratio tests for maternal and fetal genetic effects on obstetric complications. *Genet. Epidemiol.* **33** 526–538.
- DUDBRIDGE, F. (2008). Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum. Hered.* **66** 87–98. <https://doi.org/10.1159/000119108>
- FREDERICK, I. O., WILLIAMS, M. A., SALES, A. E., MARTIN, D. P. and KILLIEN, M. (2008). Pre-pregnancy body mass index, gestational weight gain, and other maternal characteristics in relation to infant birth weight. *Matern. Child Health J.* **12** 557–567.
- GJERDEVIK, M., JUGESSUR, A., HAALAND, Ø. A., ROMANOWSKA, J., LIE, R. T., CORDELL, H. J. and GJESSING, H. K. (2019). Haplin power analysis: A software module for power and sample size calculations in genetic association analyses of family triads and unrelated controls. *BMC Bioinform.* **20** 165. <https://doi.org/10.1186/s12859-019-2727-3>
- GJESSING, H. K. and LIE, R. T. (2006). Case-parent triads: Estimating single-and double-dose effects of fetal and maternal disease gene haplotypes. *Ann. Hum. Genet.* **70** 382–396.
- GODDARD, K. A., TROMP, G., ROMERO, R., OLSON, J. M., LU, Q., XU, Z., PARIMI, N., NIEN, J. K., GOMEZ, R. et al. (2007). Candidate-gene association study of mothers with pre-eclampsia, and their infants, analyzing 775 SNPs in 190 genes. *Hum. Hered.* **63** 1–16.
- HARLAP, S., DAVIES, A. M., DEUTSCH, L., CALDERON-MARGALIT, R., MANOR, O., PALTIEL, O., TIRAM, E., YANETZ, R., PERRIN, M. C. et al. (2007). The Jerusalem perinatal study cohort, 1964–2005: Methods and a review of the main results. *Paediatr. Perinat. Epidemiol.* **21** 256–273.
- HOCHNER, H., FRIEDLANDER, Y., CALDERON-MARGALIT, R., MEINER, V., SAGY, Y., AVGIL-TSADOK, M., BURGER, A., SAVITSKY, B., SISCOVICK, D. S. et al. (2012). Associations of maternal prepregnancy body mass index and gestational weight gain with adult offspring cardiometabolic risk factors: The Jerusalem perinatal family follow-up study. *Circulation* **125** 1381–1389.
- HOWEY, R. and CORDELL, H. J. (2012). PREMIM and EMIM: Tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. *BMC Bioinform.* **13** 149. <https://doi.org/10.1186/1471-2105-13-149>
- HU, G., TIAN, H., ZHANG, F., LIU, H., ZHANG, C., ZHANG, S., WANG, L., LIU, G., YU, Z. et al. (2012). Tianjin gestational diabetes mellitus prevention program: Study design, methods, and 1-year interim report on the feasibility of lifestyle intervention program. *Diabetes Res. Clin. Pract.* **98** 508–517.
- INFANTERIVARD, C. (2007). Studying genetic predisposition among small-for-gestational-age newborns. *Semin. Perinatol.* **31** 213–218.
- KANAYAMA, N., TAKAHASHI, K., MATSUURA, T., SUGIMURA, M., KOBAYASHI, T., MONIWA, N., TOMITA, M. and NAKAYAMA, K. (2002). Deficiency in p57Kip2 expression induces preeclampsia-like symptoms in mice. *Mol. Hum. Reprod.* **8** 1129–1135.
- KISTNER, E. O., SHI, M. and WEINBERG, C. R. (2009). Using cases and parents to study multiplicative gene-by-environment interaction. *Am. J. Epidemiol.* **170** 393–400. <https://doi.org/10.1093/aje/kwp118>
- KRAFT, P., YEN, Y. C., STRAM, D. O., MORRISON, J. and GAUDERMAN, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Hum. Hered.* **63** 111–119.
- LUNDE, A., MELVE, K. K., GJESSING, H. K., SKJAERVEN, R. and IRGENS, L. M. (2007). Genetic and environmental influences on birth weight, birth length, head circumference, and gestational age by use of population-based parent-offspring data. *Am. J. Epidemiol.* **167** 734–741.
- MALLIA, T., GRECH, A., HILI, A., CALLEJA-AGIUS, J. and PACE, N. P. (2017). Genetic determinants of low birth weight. *Minerva Ginecol.* **69** 631–643. <https://doi.org/10.23736/S0026-4784.17.04050-3>
- PETRY, C. J., ONG, K. K. and DUNGER, D. B. (2007). Does the fetal genotype affect maternal physiology during pregnancy. *Trends Mol. Med.* **13** 414–421.
- PIEGORSCH, W. W., WEINBERG, C. R. and TAYLOR, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat. Med.* **13** 153–162.
- PRENTICE, R. L. and PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–411. <https://doi.org/10.1093/biomet/66.3.403>
- RADHA, V., KANTHIMATHI, S., ANJANA, R. M. and MOHAN, V. (2016). Genetics of gestational diabetes mellitus. *J. Pak. Med. Assoc.* **9** Suppl 1 S11–14.
- SAFTLAS, A. F., BEYDOUN, H. and TRICHE, E. (2005). Immunogenetic determinants of preeclampsia and related pregnancy disorders: A systematic review. *Obstet. Gynecol.* **106** 162–172.
- SHI, M., UMBACH, D. M. and WEINBERG, C. R. (2011). Family-based gene-by-environment interaction studies: Revelations and remedies. *Epidemiology* **22** 400–407.
- SHI, M., UMBACH, D. M., VERMEULEN, S. H. and WEINBERG, C. R. (2008). Making the most of case-mother/control-mother studies. *Am. J. Epidemiol.* **168** 541–547.
- SKARE, Ø., JUGESSUR, A., LIE, R. T., WILCOX, A. J., MURRAY, J. C., LUNDE, A., NGUYEN, T. T. and GJESSING, H. K. (2012). Application of a novel hybrid study design to explore gene-environment interactions in orofacial clefts. *Ann. Hum. Genet.* **76** 221–236.

- UMBACH, D. and WEINBERG, C. (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat. Med.* **16** 1731–1743.
- UMBACH, D. M. and WEINBERG, C. R. (2000). The use of case-parent triads to study joint effects of genotype and exposure. *Am. J. Hum. Genet.* **66** 251–261.
- WANGLER, M. F., CHANG, A. S., MOLEY, K. H., FEINBERG, A. P. and DEBAUN, M. R. (2005). Factors associated with preterm delivery in mothers of children with Beckwith–Wiedemann syndrome: A case cohort study from the BWS registry. *Am. J. Med. Genet., Part A* **134** 187–191.
- WEINBERG, C. R. and UMBACH, D. M. (2005). A hybrid design for studying genetic influences on risk of diseases with onset early in life. *Am. J. Hum. Genet.* **77** 627–636.
- WEINBERG, C. R., WILCOX, A. J. and LIE, R. T. (1998). A log-linear approach to case-parent-triad data: Assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am. J. Hum. Genet.* **62** 969–978. <https://doi.org/10.1086/301802>
- WILCOX, A. J., WEINBERG, C. R. and LIE, R. T. (1998). Distinguishing the effects of maternal and offspring genes through studies of “case-parent triads”. *Am. J. Epidemiol.* **148** 893–901. <https://doi.org/10.1093/oxfordjournals.aje.a009715>
- YANG, J. and LIN, S. (2013). Robust partial likelihood approach for detecting imprinting and maternal effects using case-control families. *Ann. Appl. Stat.* **7** 249–268. [MR3086418 https://doi.org/10.1214/12-AOAS577](https://doi.org/10.1214/12-AOAS577)
- ZHANG, H., CHATTERJEE, N., RADER, D. and CHEN, J. (2018). Adjustment of nonconfounding covariates in case-control genetic association studies. *Ann. Appl. Stat.* **12** 200–221. [MR3773391 https://doi.org/10.1214/17-AOAS1065](https://doi.org/10.1214/17-AOAS1065)
- ZHANG, H., MUKHERJEE, B., ARTHUR, V., HU, G., HOCHNER, H. and CHEN, J. (2020). Supplement to “An efficient and computationally robust statistical method for analyzing case-control mother–offspring pair genetic association studies.” <https://doi.org/10.1214/19-AOAS1298SUPP>.

BASELINE DRIFT ESTIMATION FOR AIR QUALITY DATA USING QUANTILE TREND FILTERING

BY HALLEY L. BRANTLEY^{1,*}, JOSEPH GUINNESS² AND ERIC C. CHI^{1,†}

¹Department of Statistics, North Carolina State University, *hlbrantl@ncsu.edu; †eric_chi@ncsu.edu

²Department of Statistics and Data Science, Cornell University, guinness@cornell.edu

We address the problem of estimating smoothly varying baseline trends in time series data. This problem arises in a wide range of fields, including chemistry, macroeconomics and medicine; however, our study is motivated by the analysis of data from low cost air quality sensors. Our methods extend the quantile trend filtering framework to enable the estimation of multiple quantile trends simultaneously while ensuring that the quantiles do not cross. To handle the computational challenge posed by very long time series, we propose a parallelizable alternating direction method of multipliers (ADMM) algorithm. The ADMM algorithm enables the estimation of trends in a piecewise manner, both reducing the computation time and extending the limits of the method to larger data sizes. We also address smoothing parameter selection and propose a modified criterion based on the extended Bayesian information criterion. Through simulation studies and our motivating application to low cost air quality sensor data, we demonstrate that our model provides better quantile trend estimates than existing methods and improves signal classification of low-cost air quality sensor output.

REFERENCES

- 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.
- BRANTLEY, H. L., GUINNESS, J. and CHI, E. C. (2020). Supplement to “Baseline drift estimation for air quality data using quantile trend filtering.” <https://doi.org/10.1214/19-AOAS1318SUPPA>, <https://doi.org/10.1214/19-AOAS1318SUPPB>.
- BRANTLEY, H., HAGLER, G., KIMBROUGH, E., WILLIAMS, R., MUKERJEE, S. and NEAS, L. (2014). Mobile air monitoring data-processing strategies and effects on spatial air pollution trends. *Atmos. Meas. Tech.* **7** 2169–2183.
- 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](https://doi.org/10.1093/biomet/asn034)
- COOMBES, K. R., BAGGERLY, K. A. and MORRIS, J. S. (2007). Pre-processing mass spectrometry data. In *Fundamentals of Data Mining in Genomics and Proteomics* 79–102. Springer, Berlin.
- DU, P., KIBBE, W. A. and LIN, S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* **22** 2059–2065.
- FAULKNER, J. R. and MININ, V. N. (2018). Locally adaptive smoothing with Markov random fields and shrinkage priors. *Bayesian Anal.* **13** 225–252. [MR3737950 https://doi.org/10.1214/17-BA1050](https://doi.org/10.1214/17-BA1050)
- FRANCO, M. A. D. M., MILORI, D. M. B. P. and BOAS, P. R. V. (2018). Comparison of algorithms for baseline correction of LIBS spectra for quantifying total carbon in Brazilian soils. ArXiv Preprint. Available at [arXiv:1805.03695](https://arxiv.org/abs/1805.03695).
- GABAY, D. and MERCIER, B. (1975). A Dual Algorithm for the Solution of Non Linear Variational Problems via Finite Element Approximation. Institut de recherche d’informatique et d’automatique.
- GLOWINSKI, R. and MARROCCO, A. (1975). Sur l’approximation, par éléments finis d’ordre un, et la résolution, par pénalisation-dualité, d’une classe de problèmes de Dirichlet non linéaires. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér.* **9** 41–76. [MR0388811](https://doi.org/10.1016/S0378-4222(75)80001-1)
- GUROBI OPTIMIZATION, LLC (2018). Gurobi Optimizer Reference Manual.
- KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. [MR2505584 https://doi.org/10.1137/070690274](https://doi.org/10.1137/070690274)

- KOENKER, R. (2018). quantreg: Quantile Regression. R package version 5.36.
- KOENKER, R. and BASSETT, G. JR. (1978). Regression quantiles. *Econometrica* **46** 33–50. [MR0474644 https://doi.org/10.2307/1913643](https://doi.org/10.2307/1913643)
- KOENKER, R., NG, P. and PORTNOY, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673–680. [MR1326417 https://doi.org/10.1093/biomet/81.4.673](https://doi.org/10.1093/biomet/81.4.673)
- LUO, Y., HARGRAVES, R. H., BELLE, A., BAI, O., QI, X., WARD, K. R., PFAFFENBERGER, M. P. and NAJARIAN, K. (2013). A hierarchical method for removal of baseline drift from biomedical signals: Application in ECG analysis. *Sci. World J.* **2013**.
- MARANDI, R. Z. and SABZPOUSHAN, S. (2015). Qualitative modeling of the decision-making process using electrooculography. *Behav. Res. Methods* **47** 1404–1412.
- MECOZZI, M. (2014). A polynomial curve fitting method for baseline drift correction in the chromatographic analysis of hydrocarbons in environmental samples. *APCBEE Proc.* **10** 2–6.
- NING, X., SELESNICK, I. W. and DUVAL, L. (2014). Chromatogram baseline estimation and denoising using sparsity (BEADS). *Chemom. Intell. Lab. Syst.* **139** 156–167.
- NYCHKA, D., GRAY, G., HAALAND, P., MARTIN, D. and O'CONNELL, M. (1995). A nonparametric regression approach to syringe grading for quality improvement. *J. Amer. Statist. Assoc.* **90** 1171–1178.
- NYCHKA, D., FURRER, R., PAIGE, J. and SAIN, S. (2017). fields: Tools for Spatial Data. R package version 9.6.
- OH, H.-S., LEE, T. C. M. and NYCHKA, D. W. (2011). Fast nonparametric quantile regression with arbitrary smoothing methods. *J. Comput. Graph. Statist.* **20** 510–526. [MR2847806 https://doi.org/10.1198/jcgs.2010.10063](https://doi.org/10.1198/jcgs.2010.10063)
- OH, H.-S., NYCHKA, D., BROWN, T. and CHARBONNEAU, P. (2004). Period analysis of variable stars by robust smoothing. *J. Roy. Statist. Soc. Ser. C* **53** 15–30. [MR2037881 https://doi.org/10.1111/j.1467-9876.2004.00423.x](https://doi.org/10.1111/j.1467-9876.2004.00423.x)
- PETTERSSON, K., JAGADEESAN, S., LUKANDER, K., HENELIUS, A., HÆGGSTRÖM, E. and MÜLLER, K. (2013). Algorithm for automatic analysis of electro-oculographic data. *Biomed. Eng. Online* **12** 110.
- RACINE, J. S. and LI, K. (2017). Nonparametric conditional quantile estimation: A locally weighted quantile kernel approach. *J. Econometrics* **201** 72–94. [MR3704677 https://doi.org/10.1016/j.jeconom.2017.06.020](https://doi.org/10.1016/j.jeconom.2017.06.020)
- RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms *Phys. D, Nonlinear Phenom.* **60** 259–268.
- SNYDER, E., WATKINS, T., SOLOMON, P., THOMA, E., WILLIAMS, R., HAGLER, G., SHELOW, D., HINDIN, D., KILARU, V. et al. (2013). The changing paradigm of air pollution monitoring *Environ. Sci. Technol.* **47** 11369.
- THEUSSL, S. and HORNIK, K. (2017). Rglpk: R/GNU Linear Programming Kit Interface. R package version 0.6-3.
- THOMA, E. D., BRANTLEY, H. L., OLIVER, K. D., WHITAKER, D. A., MUKERJEE, S., MITCHELL, B., WU, T., SQUIER, B., ESCOBAR, E. et al. (2016). South Philadelphia passive sampler and sensor study. *J. Air Waste Manage. Assoc.* **66** 959–970.
- TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323. [MR3189487 https://doi.org/10.1214/13-AOS1189](https://doi.org/10.1214/13-AOS1189)
- TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. [MR2136641 https://doi.org/10.1111/j.1467-9868.2005.00490.x](https://doi.org/10.1111/j.1467-9868.2005.00490.x)
- WEINGESSEL, A. and TURLACH, B. A. (2013). quadprog: Functions to solve quadratic programming problems. R package version 1.5-5.
- WICKHAM, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.
- YAMADA, H. (2017). Estimating the trend in US real GDP using the ℓ_1 trend filtering. *Appl. Econ. Lett.* **24** 713–716.
- YU, K. and MOYEED, R. A. (2001). Bayesian quantile regression. *Statist. Probab. Lett.* **54** 437–447. [MR1861390 https://doi.org/10.1016/S0167-7152\(01\)00124-9](https://doi.org/10.1016/S0167-7152(01)00124-9)

A FAST PARTICLE-BASED APPROACH FOR CALIBRATING A 3-D MODEL OF THE ANTARCTIC ICE SHEET

BY BEN SEIYON LEE^{1,*}, MURALI HARAN^{1,†}, ROBERT W. FULLER^{2,‡}, DAVID POLLARD³ AND KLAUS KELLER^{2,§}

¹*Department of Statistics, Pennsylvania State University, *skl5261@psu.edu; †mharan@stat.psu.edu*

²*Department of Geosciences, Pennsylvania State University, ‡rwf136@psu.edu; §klaus@psu.edu*

³*Earth and Environmental Systems Institute, Pennsylvania State University, pollard@essc.psu.edu*

We consider the scientifically challenging and policy-relevant task of understanding the past and projecting the future dynamics of the Antarctic ice sheet. The Antarctic ice sheet has shown a highly nonlinear threshold response to past climate forcings. Triggering such a threshold response through anthropogenic greenhouse gas emissions would drive drastic and potentially fast sea level rise with important implications for coastal flood risks. Previous studies have combined information from ice sheet models and observations to calibrate model parameters. These studies have broken important new ground but have either adopted simple ice sheet models or have limited the number of parameters to allow for the use of more complex models. These limitations are largely due to the computational challenges posed by calibration as models become more computationally intensive or when the number of parameters increases.

Here, we propose a method to alleviate this problem: a fast sequential Monte Carlo method that takes advantage of the massive parallelization afforded by modern high-performance computing systems. We use simulated examples to demonstrate how our sample-based approach provides accurate approximations to the posterior distributions of the calibrated parameters. The drastic reduction in computational times enables us to provide new insights into important scientific questions, for example, the impact of Pliocene era data and prior parameter information on sea level projections. These studies would be computationally prohibitive with other computational approaches for calibration such as Markov chain Monte Carlo or emulation-based methods. We also find considerable differences in the distributions of sea level projections when we account for a larger number of uncertain parameters. For example, based on the same ice sheet model and data set, the 99th percentile of the Antarctic ice sheet contribution to sea level rise in 2300 increases from 6.5 m to 13.1 m when we increase the number of calibrated parameters from three to 11. With previous calibration methods, it would be challenging to go beyond five parameters. This work provides an important next step toward improving the uncertainty quantification of complex, computationally intensive and decision-relevant models.

REFERENCES

- AGAPIOU, S., PAPASPILIOPOULOS, O., SANZ-ALONSO, D. and STUART, A. M. (2017). Importance sampling: Intrinsic dimension and computational cost. *Statist. Sci.* **32** 405–431. [MR3696003 https://doi.org/10.1214/17-STS611](https://doi.org/10.1214/17-STS611)
- ALDER, J. R., HOSTETLER, S. W., POLLARD, D. and SCHMITTNER, A. (2011). Evaluation of a present-day climate simulation with a new coupled atmosphere-ocean model GENMOM. *Geosci. Model Dev.* **4** 69–83.
- APPLEGATE, P. J., KIRCHNER, N., STONE, E. J., KELLER, K. and GREVE, R. (2012). An assessment of key model parametric uncertainties in projections of Greenland ice sheet behavior. *Cryosphere* **6** 589–606.

- BAKKER, A. M., APPLGATE, P. J. and KELLER, K. (2016). A simple, physically motivated model of sea-level contributions from the Greenland ice sheet in response to temperature changes. *Environmental Modelling & Software* **83** 27–35.
- BALLARD, G., SIEFERT, C. and HU, J. (2016). Reducing communication costs for sparse matrix multiplication within algebraic multigrid. *SIAM J. Sci. Comput.* **38** C203–C231. [MR3507096](#) <https://doi.org/10.1137/15M1028807>
- BASTOS, L. S. and O’HAGAN, A. (2009). Diagnostics for Gaussian process emulators. *Technometrics* **51** 425–438. [MR2756478](#) <https://doi.org/10.1198/TECH.2009.08019>
- BAYARRI, M. J., BERGER, J. O., CAFFEO, J., GARCIA-DONATO, G., LIU, F., PALOMO, J., PARTHASARATHY, R. J., PAULO, R., SACKS, J. et al. (2007). Computer model validation with functional output. *Ann. Statist.* **35** 1874–1906. [MR2363956](#) <https://doi.org/10.1214/009053607000000163>
- BESKOS, A., JASRA, A., KANTAS, N. and THIERY, A. (2016). On the convergence of adaptive sequential Monte Carlo methods. *Ann. Appl. Probab.* **26** 1111–1146. [MR3476634](#) <https://doi.org/10.1214/15-AAP1113>
- BHAT, K. S., HARAN, M., GOES, M. and CHEN, M. (2010). Computer model calibration with multivariate spatial output: A case study. *Frontiers of Statistical Decision Making and Bayesian Analysis* 168–184.
- BHATTACHARYYA, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā* **7** 401–406. [MR0018387](#)
- BONY, S. and EMANUEL, K. A. (2001). A parameterization of the cloudiness associated with cumulus convection; evaluation using TOGA COARE data. *J. Atmos. Sci.* **58** 3158–3183.
- BRYNJARSDÓTTIR, J. and O’HAGAN, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Probl.* **30** 114007, 24. [MR3274591](#) <https://doi.org/10.1088/0266-5611/30/11/114007>
- CHANG, W., HARAN, M., OLSON, R. and KELLER, K. (2014). Fast dimension-reduced climate model calibration and the effect of data aggregation. *Ann. Appl. Stat.* **8** 649–673. [MR3262529](#) <https://doi.org/10.1214/14-AOAS733>
- CHANG, W., HARAN, M., APPLGATE, P. and POLLARD, D. (2016a). Calibrating an ice sheet model using high-dimensional binary spatial data. *J. Amer. Statist. Assoc.* **111** 57–72. [MR3494638](#) <https://doi.org/10.1080/01621459.2015.1108199>
- CHANG, W., HARAN, M., APPLGATE, P. and POLLARD, D. (2016b). Improving ice sheet model calibration using paleoclimate and modern data. *Ann. Appl. Stat.* **10** 2274–2302. [MR3592057](#) <https://doi.org/10.1214/16-AOAS979>
- CHOPIN, N. (2002). A sequential particle filter method for static models. *Biometrika* **89** 539–551. [MR1929161](#) <https://doi.org/10.1093/biomet/89.3.539>
- COLLINS, M. (2007). Ensembles and probabilities: A new era in the prediction of climate change. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **365** 1957–1970.
- COMPUTATIONAL AND INFORMATION SYSTEMS LABORATORY (2017). Cheyenne: HPE/SGI ICE XA System (University Community Computing). Boulder, CO: National Center for Atmospheric Research. doi:10.5065/D6RX99HX. <https://doi.org/10.5065/D6RX99HX>
- COOK, C. P., VAN DE FLIERDT, T., WILLIAMS, T., HEMMING, S. R., IWAI, M., KOBAYASHI, M., JIMENEZ-ESPEJO, F. J., ESCUTIA, C., GONZÁLEZ, J. J. et al. (2013). Dynamic behaviour of the East Antarctic ice sheet during Pliocene warmth. *Nature Geoscience* **6** 765–769.
- COOK, C., HILL, D., VAN DE FLIERDT, T., WILLIAMS, T., HEMMING, S., DOLAN, A., PIERCE, E., ESCUTIA, C., HARWOOD, D. et al. (2014). Sea surface temperature control on the distribution of far-traveled Southern Ocean ice-rafted detritus during the Pliocene. *Paleoceanography* **29** 533–548.
- CRISAN, D. and DOUCET, A. (2000). Convergence of sequential Monte Carlo methods. Technical Report CUED/F-INFENG/TR381, Signal Processing Group, Department of Engineering, University of Cambridge.
- DAL GESSO, S. and NEGGERS, R. (2018). Can we use single-column models for understanding the boundary layer cloud-climate feedback? *J. Adv. Model. Earth Syst.* **10** 245–261.
- DE BOER, B., STOCCHI, P. and VAN DE WAL, R. (2014). A fully coupled 3-D ice-sheet-sea-level model: Algorithm and applications. *Geosci. Model Dev.* **7** 2141–2156.
- DECONTO, R. M. and POLLARD, D. (2016). Contribution of Antarctica to past and future sea-level rise. *Nature* **531** 591–597. <https://doi.org/10.1038/nature17145>
- DEL MORAL, P., DOUCET, A. and JASRA, A. (2006). Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 411–436. [MR2278333](#) <https://doi.org/10.1111/j.1467-9868.2006.00553.x>
- DESCHAMPS, P., DURAND, N., BARD, E., HAMELIN, B., CAMOIN, G., THOMAS, A. L., HENDERSON, G. M., OKUNO, J. and YOKOYAMA, Y. (2012). Ice-sheet collapse and sea-level rise at the Bølling warming 14,600 years ago. *Nature* **483** 559.
- DIAZ, D. and KELLER, K. (2016). A potential disintegration of the West Antarctic ice sheet: implications for economic analyses of climate policy. *Am. Econ. Rev.* **106** 607–11.

- DOLAN, A. M., HAYWOOD, A. M., HILL, D. J., DOWSETT, H. J., HUNTER, S. J., LUNT, D. J. and PICKERING, S. J. (2011). Sensitivity of Pliocene ice sheets to orbital forcing. *Palaeogeography, Palaeoclimatology, Palaeoecology* **309** 98–110.
- DOLAN, A. M., DE BOER, B., BERNALES, J., HILL, D. J. and HAYWOOD, A. M. (2018). High climate model dependency of Pliocene Antarctic ice-sheet predictions. *Nat. Commun.* **9** 2799. <https://doi.org/10.1038/s41467-018-05179-4>
- DOUCET, A., DE FREITAS, N. and GORDON, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo Methods in Practice*. *Stat. Eng. Inf. Sci.* 3–14. Springer, New York. MR1847784 https://doi.org/10.1007/978-1-4757-3437-9_1
- DOUCET, A., GODSILL, S. and ANDRIEU, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comput.* **10** 197–208.
- DOWSETT, H. J. and CRONIN, T. M. (1990). High eustatic sea level during the middle Pliocene: evidence from the southeastern US Atlantic Coastal Plain. *Geology* **18** 435–438.
- DUTTON, A., CARLSON, A. E., LONG, A. J., MILNE, G. A., CLARK, P. U., DECONTO, R., HORTON, B. P., RAHMSTORF, S. and RAYMO, M. E. (2015). SEA-LEVEL RISE. Sea-level rise due to polar ice-sheet mass loss during past warm periods. *Science* **349** aaa4019. <https://doi.org/10.1126/science.aaa4019>
- EDWARDS, T. L., BRANDON, M. A., DURAND, G., EDWARDS, N. R., GOLLEDGE, N. R., HOLDEN, P. B., NIAS, I. J., PAYNE, A. J., RITZ, C. et al. (2019). Revisiting Antarctic ice loss due to marine ice-cliff instability. *Nature* **566** 58–64. <https://doi.org/10.1038/s41586-019-0901-4>
- FAN, W., YU, W., XU, J., ZHOU, J., LUO, X., YIN, Q., LU, P., CAO, Y. and XU, R. (2018). Parallelizing sequential graph computations. *ACM Trans. Database Syst.* **43** Art. 18, 39. MR3892830 <https://doi.org/10.1145/3282488>
- FITZGERALD, P., BAMBER, J., RIDLEY, J. and ROUGIER, J. (2012). Exploration of parametric uncertainty in a surface mass balance model applied to the Greenland ice sheet. *Journal of Geophysical Research: Earth Surface* **117**.
- FRETWELL, P., PRITCHARD, H., VAUGHAN, D., BAMBER, J., BARRAND, N., BELL, R., BIANCHI, C., BINGHAM, R., BLANKENSHIP, D. et al. (2012). Bedmap2: improved ice bed, surface and thickness datasets for Antarctica. *The Cryosphere Discussions* **6** 4305–4361.
- FULLER, R. W., WONG, T. E. and KELLER, K. (2017). Probabilistic inversion of expert assessments to inform projections about Antarctic ice sheet responses. *PLoS ONE* **12** e0190115. <https://doi.org/10.1371/journal.pone.0190115>
- GARNER, G. G. and KELLER, K. (2018). Using direct policy search to identify robust strategies in adapting to uncertain sea-level rise and storm surge. *Environmental Modelling & Software* **107** 96–104.
- GETTELMAN, A., TRUESDALE, J. E., BACMEISTER, J. T., CALDWELL, P. M., NEALE, R. B., BOGENSCHUTZ, P. A. and SIMPSON, I. R. (2019). The single column atmosphere model version 6 (SCAM6): Not a scam but a tool for model evaluation and development. *J. Adv. Model. Earth Syst.* **0**.
- GILKS, W. R. and BERZUINI, C. (2001). Following a moving target—Monte Carlo inference for dynamic Bayesian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 127–146. MR1811995 <https://doi.org/10.1111/1467-9868.00280>
- GIRAUD, F. and DEL MORAL, P. (2017). Nonasymptotic analysis of adaptive and annealed Feynman–Kac particle models. *Bernoulli* **23** 670–709. MR3556789 <https://doi.org/10.3150/14-BEJ680>
- GOLLEDGE, N. R., KELLER, E. D., GOMEZ, N., NAUGHTEN, K. A., BERNALES, J., TRUSEL, L. D. and EDWARDS, T. L. (2019). Global environmental consequences of twenty-first-century ice-sheet melt. *Nature* **566** 65–72. <https://doi.org/10.1038/s41586-019-0889-9>
- GOODMAN, J. and WEARE, J. (2010). Ensemble samplers with affine invariance. *Commun. Appl. Math. Comput. Sci.* **5** 65–80. MR2600822 <https://doi.org/10.2140/camcos.2010.5.65>
- GORDON, N. J., SALMOND, D. J. and SMITH, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEEE Proceedings F-Radar and Signal Processing* **140** 107–113. IET.
- GRAMACY, R. B. and APLEY, D. W. (2015). Local Gaussian process approximation for large computer experiments. *J. Comput. Graph. Statist.* **24** 561–578. MR3357395 <https://doi.org/10.1080/10618600.2014.914442>
- GREVE, R. (1997). Application of a polythermal three-dimensional ice sheet model to the Greenland ice sheet: Response to steady-state and transient climate scenarios. *J. Climate* **10** 901–918.
- HANNART, A., GHIL, M., DUFRESNE, J.-L. and NAVEAU, P. (2013). Disconcerting learning on climate sensitivity and the uncertain future of uncertainty. *Clim. Change* **119** 585–601.
- HIGDON, D., GATTIKER, J., WILLIAMS, B. and RIGHTLEY, M. (2008). Computer model calibration using high-dimensional output. *J. Amer. Statist. Assoc.* **103** 570–583. MR2523994 <https://doi.org/10.1198/016214507000000888>
- ISAAC, T., STADLER, G. and GHATTAS, O. (2015). Solution of nonlinear Stokes equations discretized by high-order finite elements on nonconforming and anisotropic meshes, with application to ice sheet dynamics. *SIAM J. Sci. Comput.* **37** B804–B833. MR3418230 <https://doi.org/10.1137/140974407>

- ISAAC, T., PETRA, N., STADLER, G. and GHATTAS, O. (2015). Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the Antarctic ice sheet. *J. Comput. Phys.* **296** 348–368. [MR3348506](#) <https://doi.org/10.1016/j.jcp.2015.04.047>
- JACKSON, C. H., JIT, M., SHARPLES, L. D. and DE ANGELIS, D. (2015). Calibration of complex models through Bayesian evidence synthesis: A demonstration and tutorial. *Medical Decision Making* **35** 148–161.
- JASRA, A., STEPHENS, D. A., DOUCET, A. and TSAGARIS, T. (2011). Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. *Scand. J. Stat.* **38** 1–22. [MR2760137](#) <https://doi.org/10.1111/j.1467-9469.2010.00723.x>
- JEREMIAH, E., SISSON, S., MARSHALL, L., MEHROTRA, R. and SHARMA, A. (2011). Bayesian calibration and uncertainty analysis of hydrological models: A comparison of adaptive Metropolis and sequential Monte Carlo samplers. *Water Resour. Res.* **47**.
- JOHNSON, D. R., FISCHBACH, J. R. and ORTIZ, D. S. (2013). Estimating surge-based flood risk with the coastal Louisiana risk assessment model. *Journal of Coastal Research* **67** 109–126.
- JONES, G. L., HARAN, M., CAFFO, B. S. and NEATH, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **101** 1537–1547. [MR2279478](#) <https://doi.org/10.1198/016214506000000492>
- KALYANARAMAN, J., KAWAJIRI, Y., LIVELY, R. P. and REALFF, M. J. (2016). Uncertainty quantification via Bayesian inference using sequential Monte Carlo methods for CO₂ adsorption process. *AIChE Journal* **62** 3352–3368.
- KANTAS, N., BESKOS, A. and JASRA, A. (2014). Sequential Monte Carlo methods for high-dimensional inverse problems: A case study for the Navier–Stokes equations. *SIAM/ASA J. Uncertain. Quantificat.* **2** 464–489. [MR3283917](#) <https://doi.org/10.1137/130930364>
- KANTAS, N., DOUCET, A., SINGH, S. S., MACIEJOWSKI, J. and CHOPIN, N. (2015). On particle methods for parameter estimation in state-space models. *Statist. Sci.* **30** 328–351. [MR3383884](#) <https://doi.org/10.1214/14-STS511>
- KELLER, K. and MCINERNEY, D. (2008). The dynamics of learning about a climate threshold. *Clim. Dyn.* **30** 321–332.
- KENNEDY, M. C. and O’HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. [MR1858398](#) <https://doi.org/10.1111/1467-9868.00294>
- KIM, S. H., EDMONDS, J., LURZ, J., SMITH, S. J. and WISE, M. (2006). The objECTS framework for integrated assessment: Hybrid modeling of transportation. *Energy J.* 63–91.
- KONG, A. (1992). A note on importance sampling using standardized weights. University of Chicago, Dept. of Statistics, Tech. Rep, 348.
- KOPP, R. E., SIMONS, F. J., MITROVICA, J. X., MALOOF, A. C. and OPPENHEIMER, M. (2009). Probabilistic assessment of sea level during the last interglacial stage. *Nature* **462** 863–867. <https://doi.org/10.1038/nature08686>
- LAROUR, E., SEROUSSI, H., MORLIGHEM, M. and RIGNOT, E. (2012). Continental scale, high order, high spatial resolution, ice sheet modeling using the Ice Sheet System Model (ISSM). *Journal of Geophysical Research: Earth Surface* **117**.
- LE BARS, D., DRIJFHOUT, S. and DE VRIES, H. (2017). A high-end sea level rise probabilistic projection including rapid Antarctic ice sheet mass loss. *Environ. Res. Lett.* **12** 044013.
- LE BROcq, A. M., PAYNE, A. J. and VIELI, A. (2010). An improved Antarctic dataset for high resolution numerical ice sheet models (ALBMAP v1). *Earth System Science Data* **2** 247–260.
- LE COZANNET, G., MANCEAU, J.-C. and ROHMER, J. (2017). Bounding probabilistic sea-level projections within the framework of the possibility theory. *Environ. Res. Lett.* **12** 014012.
- LEE, B. S., HARAN, M., FULLER, R. W., POLLARD, D. and KELLER, K. (2020). Supplement to “A fast particle-based approach for calibrating a 3-D model of the Antarctic ice sheet.” <https://doi.org/10.1214/19-AOAS1305SUPP>.
- LELE, S. R., DENNIS, B. and LUTSCHER, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecol. Lett.* **10** 551–563.
- LEVITUS, S., ANTONOV, J. I., BOYER, T. P., BARANOVA, O. K., GARCIA, H. E., LOCARNINI, R. A., MISHONOV, A. V., REAGAN, J., SEIDOV, D. et al. (2012). World ocean heat content and thermosteric sea level change (0–2000 m). *Geophys. Res. Lett.* **39** 1955–2010.
- LI, T., SUN, S., SATTAR, T. P. and CORCHADO, J. M. (2014). Fight sample degeneracy and impoverishment in particle filters: A review of intelligent approaches. *Expert Systems with Applications* **41** 3944–3954.
- LIANG, F. and WONG, W. H. (2001). Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *J. Amer. Statist. Assoc.* **96** 653–666. [MR1946432](#) <https://doi.org/10.1198/016214501753168325>

- LIANG, X., LETTENMAIER, D. P., WOOD, E. F. and BURGESS, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res., Atmos.* **99** 14415–14428.
- LIU, F., BAYARRI, M. J. and BERGER, J. O. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Anal.* **4** 119–150. [MR2486241](#) <https://doi.org/10.1214/09-BA404>
- LIU, J. S. and CHEN, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.* **93** 1032–1044. [MR1649198](#) <https://doi.org/10.2307/2669847>
- LIU, X. and GUILLAS, S. (2017). Dimension reduction for Gaussian process emulation: An application to the influence of bathymetry on tsunami heights. *SIAM/ASA J. Uncertain. Quantificat.* **5** 787–812. [MR3686816](#) <https://doi.org/10.1137/16M1090648>
- LIU, J. S., LIANG, F. and WONG, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *J. Amer. Statist. Assoc.* **95** 121–134. [MR1803145](#) <https://doi.org/10.2307/2669532>
- LIU, J. and WEST, M. (2001). Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice*. *Stat. Eng. Inf. Sci.* 197–223. Springer, New York. [MR1847793](#)
- LIU, Z., OTTO-BLIESNER, B., HE, F., BRADY, E., TOMAS, R., CLARK, P., CARLSON, A., LYNCH-STIEGLITZ, J., CURRY, W. et al. (2009). Transient simulation of last deglaciation with a new mechanism for Bølling–Allerød warming. *Science* **325** 310–314.
- LLOPIS, F. P., KANTAS, N., BESKOS, A. and JASRA, A. (2018). Particle filtering for stochastic Navier–Stokes signal observed with linear additive noise. *SIAM J. Sci. Comput.* **40** A1544–A1565. [MR3807951](#) <https://doi.org/10.1137/17M1151900>
- LORENZ, E. N. (1972). Predictability: Does the flap of a butterfly’s wings in Brazil set off a tornado in Texas? Presented at the 139th Meeting of the AAAS. Available at http://eaps4.mit.edu/research/Lorenz/Butterfly_1972.pdf. Accessed: 2019-08-04.
- MANIYAR, D., CORNFORD, D. and BOUKOUVALAS, A. (2007). Dimensionality reduction in the emulator setting. Technical report, Neural Computing Research Group, University of Aston.
- MARTINO, L. (2018). A review of multiple try MCMC algorithms for signal processing. *Digit. Signal Process.* **75** 134–152. [MR3765591](#) <https://doi.org/10.1016/j.dsp.2018.01.004>
- MARTINO, L., ELVIRA, V. and LOUZADA, F. (2017). Effective sample size for importance sampling based on discrepancy measures. *Signal Process.* **131** 386–401.
- MCGRANAHAN, G., BALK, D. and ANDERSON, B. (2007). The rising tide: Assessing the risks of climate change and human settlements in low elevation coastal zones. *Environment and Urbanization* **19** 17–37.
- MCKAY, M. D., BECKMAN, R. J. and CONOVER, W. J. (2000). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **42** 55–61.
- MEINSHAUSEN, M., SMITH, S. J., CALVIN, K., DANIEL, J. S., KAINUMA, M., LAMARQUE, J.-F., MATSUMOTO, K., MONTZKA, S., RAPER, S. et al. (2011). The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Clim. Change* **109** 213.
- MONIER, E., SCOTT, J., SOKOLOV, A., FOREST, C. and SCHLOSSER, C. (2013). An integrated assessment modelling framework for uncertainty studies in global and regional climate change: The MIT IGSM-CAM (version 1.0). *Geoscientific Model Development Discussions* **6**.
- MORZFELD, M., TU, X., WILKENING, J. and CHORIN, A. J. (2015). Parameter estimation by implicit sampling. *Commun. Appl. Math. Comput. Sci.* **10** 205–225. [MR3402350](#) <https://doi.org/10.2140/camcos.2015.10.205>
- MORZFELD, M., DAY, M. S., GROUT, R. W., PAU, G. S. H., FINSTERLE, S. A. and BELL, J. B. (2018). Iterative importance sampling algorithms for parameter estimation. *SIAM J. Sci. Comput.* **40** B329–B352. [MR3769709](#) <https://doi.org/10.1137/16M1088417>
- MURRAY, L. M., LEE, A. and JACOB, P. E. (2016). Parallel resampling in the particle filter. *J. Comput. Graph. Statist.* **25** 789–805. [MR3533638](#) <https://doi.org/10.1080/10618600.2015.1062015>
- MURRAY, T., SELMES, N., JAMES, T. D., EDWARDS, S., MARTIN, I., O’FARRELL, T., ASPEY, R., RUTT, I., NETTLES, M. et al. (2015). Dynamics of glacier calving at the ungrounded margin of Helheim Glacier, south-east Greenland. *J. Geophys. Res. Earth Surf.* **120** 964–982. <https://doi.org/10.1002/2015JF003531>
- NAISH, T., POWELL, R., LEVY, R., WILSON, G., SCHERER, R., TALARICO, F., KRISSEK, L., NIESSEN, F., POMPILIO, M. et al. (2009). Obliquity-paced Pliocene West Antarctic ice sheet oscillations. *Nature* **458** 322–328. <https://doi.org/10.1038/nature07867>
- NEAL, R. M. (2001). Annealed importance sampling. *Stat. Comput.* **11** 125–139. [MR1837132](#) <https://doi.org/10.1023/A:1008923215028>
- NGUYEN, T. L. T. (2014). Sequential Monte-Carlo sampler for Bayesian inference in complex systems. PhD thesis, Lille 1.
- NICHOLLS, R. J., TOL, R. S. and VAFEIDIS, A. T. (2008). Global estimates of the impact of a collapse of the West Antarctic ice sheet: an application of FUND. *Clim. Change* **91** 171.
- O’NEILL, B. C., CRUTZEN, P., GRÜBLER, A., DUONG, M. H., KELLER, K., KOLSTAD, C., KOOMEY, J., LANGE, A., OBERSTEINER, M. et al. (2006). Learning and climate change. *Climate Policy* **6** 585–589.

- OEPENHEIMER, M. and ALLEY, R. B. (2016). How high will the seas rise? *Science* **354** 1375–1377. <https://doi.org/10.1126/science.aak9460>
- PAL, J. S., GIORGI, F., BI, X., ELGUINDI, N., SOLMON, F., GAO, X., RAUSCHER, S. A., FRANCISCO, R., ZAKKEY, A. et al. (2007). Regional climate modeling for the developing world: the ICTP RegCM3 and RegCNET. *Bull. Am. Meteorol. Soc.* **88** 1395–1410.
- PAPAIOANNOU, I., PAPADIMITRIOU, C. and STRAUB, D. (2016). Sequential importance sampling for structural reliability analysis. *Structural Safety* **62** 66–75.
- PETRA, N., MARTIN, J., STADLER, G. and GHATTAS, O. (2014). A computational framework for infinite-dimensional Bayesian inverse problems, Part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems. *SIAM J. Sci. Comput.* **36** A1525–A1555. MR3233941 <https://doi.org/10.1137/130934805>
- POLLARD, D. and DECONTO, R. M. (2009). Modelling West Antarctic ice sheet growth and collapse through the past five million years. *Nature* **458** 329–332. <https://doi.org/10.1038/nature07809>
- POLLARD, D. and DECONTO, R. (2012). Description of a hybrid ice sheet-shelf model, and application to Antarctica. *Geosci. Model Dev.* **5** 1273.
- POLLARD, D., DECONTO, R. M. and ALLEY, R. B. (2015). Potential Antarctic ice sheet retreat driven by hydrofracturing and ice cliff failure. *Earth and Planetary Science Letters* **412** 112–121.
- POLLARD, D., GOMEZ, N. and DECONTO, R. M. (2017). Variations of the Antarctic ice sheet in a coupled ice sheet-Earth-sea level model: Sensitivity to viscoelastic Earth properties. *Journal of Geophysical Research: Earth Surface* **122** 2124–2138.
- POLLARD, D., CHANG, W., HARAN, M., APPEGATE, P. and DECONTO, R. (2016). Large ensemble modeling of the last deglacial retreat of the West Antarctic ice sheet: comparison of simple and advanced statistical techniques. *Geosci. Model Dev.* **9** 1697–1723.
- QUEIPO, N. V., HAFTKA, R. T., SHYY, W., GOEL, T., VAIDYANATHAN, R. and TUCKER, P. K. (2005). Surrogate-based analysis and optimization. *Progress in Aerospace Sciences* **41** 1–28.
- REESE, C. S., WILSON, A. G., HAMADA, M., MARTZ, H. F. and RYAN, K. J. (2004). Integrated analysis of computer and physical experiments. *Technometrics* **46** 153–164. MR2060014 <https://doi.org/10.1198/004017004000000211>
- ROVERE, A., RAYMO, M. E., MITROVICA, J., HEARTY, P. J., O’LEARY, M. and INGLIS, J. (2014). The Mid-Pliocene sea-level conundrum: glacial isostasy, eustasy and dynamic topography. *Earth and Planetary Science Letters* **387** 27–33.
- RUCKERT, K. L., SHAFFER, G., POLLARD, D., GUAN, Y., WONG, T. E., FOREST, C. E. and KELLER, K. (2017). Assessing the impact of retreat mechanisms in a simple Antarctic ice sheet model using Bayesian calibration. *PLoS ONE* **12** e0170052.
- RUTT, I. C., HAGDORN, M., HULTON, N. and PAYNE, A. (2009). The Glimmer community ice sheet model. *Journal of Geophysical Research: Earth Surface* **114**.
- SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. With comments and a rejoinder by the authors. MR1041765
- SALZMANN, U., DOLAN, A. M., HAYWOOD, A. M., CHAN, W.-L., VOSS, J., HILL, D. J., ABE-OUCHI, A., OTTO-BLIESNER, B., BRAGG, F. J. et al. (2013). Challenges in quantifying Pliocene terrestrial warming revealed by data-model discord. *Nature Climate Change* **3** 969–974.
- SANSÓ, B., FOREST, C. E. and ZANTEDESCHI, D. (2008). Inferring climate system properties using a computer model. *Bayesian Anal.* **3** 1–37. MR2383247 <https://doi.org/10.1214/08-BA301>
- SCHÄFER, C. and CHOPIN, N. (2013). Sequential Monte Carlo on large binary sampling spaces. *Stat. Comput.* **23** 163–184. MR3016936 <https://doi.org/10.1007/s11222-011-9299-z>
- SCHLEGEL, N.-J., SEROUSSI, H., SCHODLOK, M. P., LAROUR, E. Y., BOENING, C., LIMONADI, D., WATKINS, M. M., MORLIGHEM, M. and VAN DEN BROECKE, M. R. (2018). Exploration of Antarctic ice sheet 100-year contribution to sea level rise and associated model uncertainties using the ISSM framework. *Cryosphere* **12** 3511–3534.
- SHAFFER, G. (2014). Formulation, calibration and validation of the DAIS model, a simple Antarctic ice sheet model sensitive to variations of sea level and ocean subsurface temperature. *Geosci. Model Dev.* **7** 1803–1818.
- SHIELDS, C. A. and KIEHL, J. T. (2016). Atmospheric river landfall-latitude changes in future climate simulations. *Geophys. Res. Lett.* **43** 8775–8782.
- SOROOSHIAN, S., DUAN, Q. and GUPTA, V. K. (1993). Calibration of rainfall-runoff models: Application of global optimization to the Sacramento soil moisture accounting model. *Water Resour. Res.* **29** 1185–1194.
- SRIVER, R. L., LEMPERT, R. J., WIKMAN-SVAHN, P. and KELLER, K. (2018). Characterizing uncertain sea-level rise projections to support investment decisions. *PLoS ONE* **13** e0190641. <https://doi.org/10.1371/journal.pone.0190641>
- STEIN, M. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics* **29** 143–151. MR0887702 <https://doi.org/10.2307/1269769>

- STEINBERG, D. M. and LIN, D. K. J. (2006). A construction method for orthogonal Latin hypercube designs. *Biometrika* **93** 279–288. [MR2278083](#) <https://doi.org/10.1093/biomet/93.2.279>
- STONE, E., LUNT, D., RUTT, I., HANNA, E. et al. (2010). Investigating the sensitivity of numerical model simulations of the modern state of the Greenland ice-sheet and its future response to climate change. *Cryosphere* **4** 397–417.
- URBAN, N. M. and FRICKER, T. E. (2010). A comparison of Latin hypercube and grid ensemble designs for the multivariate emulation of an Earth system model. *Comput. Geosci.* **36** 746–755.
- WHITELEY, N., LEE, A. and HEINE, K. (2016). On the role of interaction in sequential Monte Carlo algorithms. *Bernoulli* **22** 494–529. [MR3449791](#) <https://doi.org/10.3150/14-BEJ666>
- WILLEMS, J. C. (1972). Dissipative dynamical systems. I. General theory. *Arch. Ration. Mech. Anal.* **45** 321–351. [MR0527462](#) <https://doi.org/10.1007/BF00276493>
- WILLIAMSON, D., GOLDSTEIN, M., ALLISON, L., BLAKER, A., CHALLENGOR, P., JACKSON, L. and YAMAZAKI, K. (2013). History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Clim. Dyn.* **2** 1703–1729.
- WONG, T. E., BAKKER, A. M. and KELLER, K. (2017). Impacts of Antarctic fast dynamics on sea-level projections and coastal flood defense. *Clim. Change* **144** 347–364.
- ZHOU, Y., JOHANSEN, A. M. and ASTON, J. A. D. (2016). Toward automatic model comparison: An adaptive sequential Monte Carlo approach. *J. Comput. Graph. Statist.* **25** 701–726. [MR3533634](#) <https://doi.org/10.1080/10618600.2015.1060885>
- ZHU, G., LI, X., MA, J., WANG, Y., LIU, S., HUANG, C., ZHANG, K. and HU, X. (2018). A new moving strategy for the sequential Monte Carlo approach in optimizing the hydrological model parameters. *Adv. Water Resour.* **114** 164–179.

- FISHER, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge Univ. Press, Cambridge. [MR1251957](#) <https://doi.org/10.1017/CBO9780511564345>
- FORBES, P. G. M. and LAURITZEN, S. (2015). Linear estimating equations for exponential families with application to Gaussian linear concentration models. *Linear Algebra Appl.* **473** 261–283. [MR3338335](#) <https://doi.org/10.1016/j.laa.2014.08.015>
- FORBES, C., EVANS, M., HASTINGS, N. and PEACOCK, B. (2011). *Statistical Distributions*. Wiley, Hoboken, NJ. [MR2964192](#)
- HYVÄRINEN, A. (2005). Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **6** 695–709. [MR2249836](#)
- HYVÄRINEN, A. (2007). Some extensions of score matching. *Comput. Statist. Data Anal.* **51** 2499–2512. [MR2338984](#) <https://doi.org/10.1016/j.csda.2006.09.003>
- KASS, R. E., EDEN, U. T. and BROWN, E. N. (2014). *Analysis of Neural Data. Springer Series in Statistics*. Springer, New York. [MR3244261](#) <https://doi.org/10.1007/978-1-4614-9602-1>
- KLEIN, N., ORELLANA, J., BRINCAT, S., MILLER, E. K. and KASS, R. E. (2020a). Supplement A. Supplement to “Torus graphs for multivariate phase coupling analysis.” <https://doi.org/10.1214/19-AOAS1300SUPPA>.
- KLEIN, N., ORELLANA, J., BRINCAT, S., MILLER, E. K. and KASS, R. E. (2020b). Supplement B. Supplement code and data to “Torus graphs for multivariate phase coupling analysis.” <https://doi.org/10.1214/19-AOAS1300SUPPB>.
- KURZ, G. and HANEBECK, U. D. (2015). Toroidal information fusion based on the bivariate von Mises distribution. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems MFI*. IEEE.
- LACHAUX, J. P., RODRIGUEZ, E., MARTINERIE, J. and VARELA, F. J. (1999). Measuring phase synchrony in brain signals. *Hum. Brain Mapp.* **8** 194–208. [https://doi.org/10.1002/\(sici\)1097-0193\(1999\)8:4<194::aid-hbm4>3.0.co;2-c](https://doi.org/10.1002/(sici)1097-0193(1999)8:4<194::aid-hbm4>3.0.co;2-c)
- LIN, L., DRTON, M. and SHOJAIE, A. (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.* **10** 806–854. [MR3486418](#) <https://doi.org/10.1214/16-EJS1126>
- MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics. Wiley Series in Probability and Statistics*. Wiley, Chichester. [MR1828667](#)
- MARDIA, K. V., KENT, J. T. and LAHA, A. K. (2016). Score matching estimators for directional distributions. Preprint. Available at [arXiv:1604.08470](https://arxiv.org/abs/1604.08470).
- MARDIA, K. V. and PATRANGENARU, V. (2005). Directions and projective shapes. *Ann. Statist.* **33** 1666–1699. [MR2166559](#) <https://doi.org/10.1214/009053605000000273>
- MARDIA, K. V., TAYLOR, C. C. and SUBRAMANIAM, G. K. (2007). Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics* **63** 505–512. [MR2370809](#) <https://doi.org/10.1111/j.1541-0420.2006.00682.x>
- MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. [MR2758523](#) <https://doi.org/10.1111/j.1467-9868.2010.00740.x>
- RANA, K. D., VAINA, L. M. and HÄMÄLÄINEN, M. S. (2013). A fast statistical significance test for baseline correction and comparative analysis in phase locking. *Front Neuroinform* **7** 3. <https://doi.org/10.3389/fninf.2013.00003>
- RODRIGUEZ-LUJAN, L., LARRAÑAGA, P. and BIELZA, C. (2017). Frobenius norm regularization for the multivariate von Mises distribution. *Int. J. Intell. Syst.* **32** 153–176.
- SHERMAN, M. A., LEE, S., LAW, R., HAEGENS, S., THORN, C. A., HÄMÄLÄINEN, M. S., MOORE, C. I. and JONES, S. R. (2016). Neural mechanisms of transient neocortical beta rhythms: Converging evidence from humans, computational modeling, monkeys, and mice. *Proc. Natl. Acad. Sci. USA* **113** E4885–E4894.
- STEINMETZ, N. A., KOCH, C., HARRIS, K. D. and CARANDINI, M. (2018). Challenges and opportunities for large-scale electrophysiology with neuropixels probes. *Curr. Opin. Neurobiol.* **50** 92–100. <https://doi.org/10.1016/j.conb.2018.01.009>
- TORT, A. B., KOMOROWSKI, R., EICHENBAUM, H. and KOPELL, N. (2010). Measuring phase-amplitude coupling between neuronal oscillations of different frequencies. *J. Neurophysiol.* **104** 1195–1210.
- WAINWRIGHT, M. J., JORDAN, M. I. et al. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.
- YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16** 3813–3847. [MR3450553](#)
- YU, S., DRTON, M. and SHOJAIE, A. (2019). Graphical models for non-negative data using generalized score matching. Preprint. Available at [arXiv:1802.06340](https://arxiv.org/abs/1802.06340).
- YU, M., KOLAR, M. and GUPTA, V. (2016). Statistical inference for pairwise graphical models using score matching. In *Advances in Neural Information Processing Systems* 2829–2837.
- ZEMEL, R. S., WILLIAMS, C. K. I. and MOZER, M. C. (1993). Directional-unit Boltzmann machines. *Adv. Neural Inf. Process. Syst.*

THE STRATIFIED MICRO-RANDOMIZED TRIAL DESIGN: SAMPLE SIZE CONSIDERATIONS FOR TESTING NESTED CAUSAL EFFECTS OF TIME-VARYING TREATMENTS

BY WALTER DEMPSEY^{1,*}, PENG LIAO^{1,**}, SANTOSH KUMAR² AND SUSAN A. MURPHY³

¹Department of Biostatistics, University of Michigan, *wdem@umich.edu; **pengliao@umich.edu

²Department of Computer Science, University of Memphis, santosh.kumar@memphis.edu

³Department of Statistics, Harvard University, samurphy@fas.harvard.edu

Technological advancements in the field of mobile devices and wearable sensors have helped overcome obstacles in the delivery of care, making it possible to deliver behavioral treatments anytime and anywhere. Here, we discuss our work on the design of a mobile health smoking cessation intervention study with the goal of assessing whether reminders, delivered at times of stress, result in a reduction/prevention of stress in the near-term and whether this effect changes with time in study. Multiple statistical challenges arose in this effort, leading to the development of the *stratified micro-randomized trial* design. In these designs each individual is randomized to treatment repeatedly at times determined by predictions of risk. These *risk times* may be impacted by prior treatment. We describe the statistical challenges and detail how they can be met.

REFERENCES

- BIDARGADDI, N., ALMIRALL, D., MURPHY, S., NAHUM-SHANI, I., KOVALCIK, M., PITUCH, T., MAAIEH, H. and STRECHER, V. (2018). To prompt or not to prompt? A microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. *JMIR Mhealth Uhealth* **6** e10123. <https://doi.org/10.2196/10123>.
- BOJINOV, I. and SHEPHARD, N. (2019). Time series experiments and causal estimands: Exact randomization tests and trading. *J. Amer. Statist. Assoc.* **114** 1665–1682. <https://doi.org/10.1080/01621459.2018.1527225>
- BORUVKA, A., ALMIRALL, D., WITKIEWITZ, K. and MURPHY, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *J. Amer. Statist. Assoc.* **113** 1112–1121. MR3862343 <https://doi.org/10.1080/01621459.2017.1305274>
- DEMPSEY, W., LIAO, P., KLASNJA, P., NAHUN-SHANI, I. and MURPHY, S. A. (2015). Randomised trials for the fitbit generation. *Significance* **12** 20–23.
- DEMPSEY, W., LIAO, P., KUMAR, S. and MURPHY, S. A. (2020). Supplement to “The stratified micro-randomized trial design: Sample size considerations for testing nested causal effects of time-varying treatments.” <https://doi.org/10.1214/19-AOAS1293SUPP>.
- DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. Oxford Statistical Science Series **25**. Oxford Univ. Press, Oxford. MR2049007
- ERTIN, E., STOHS, N., KUMAR, S., RAIJ, A., AL’ABSI, M. and AUTONSENSE, S. S. (2011). Unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems* 274–287, New York, NY, USA.
- FREE, C., PHILLIPS, G., GALLI, L., WATSON, L., FELIX, L., EDWARDS, P., PATEL, V. and HAINES, A. (2013). The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: A systematic review. *PLoS Med.* **10** 1–45.
- HONG, G. and RAUDENBUSH, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J. Amer. Statist. Assoc.* **101** 901–910. MR2324091 <https://doi.org/10.1198/016214506000000447>
- HOTELLING, H. (1931). The generalization of student’s ratio. *Annals of Mathematical Sciences* **2** 360–378.

Key words and phrases. Sequential randomization, nested causal effects, stratified micro-randomized trials, mobile health, weighted-centered least-squares method.

- HOVSEPIAN, K., AL'ABSI, M., ERTIN, E., KAMARCK, T., NAKAJIMA, M. and KUMAR, S. (2015). cstress: Towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15* 493–504. ACM, New York, NY, USA.
- IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. [MR3309951](#) <https://doi.org/10.1017/CBO9781139025751>
- KLASNJA, P., HEKLER, E. B., SHIFFMAN, S., BORUVKA, A., ALMIRALL, D., TEWARI, A. and MURPHY, S. A. (2015). Micro-randomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology* **34** 1220–1228.
- LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. [MR0836430](#) <https://doi.org/10.1093/biomet/73.1.13>
- LIAO, P., KLASNJA, P., TEWARI, A. and MURPHY, S. A. (2016). Micro-randomized trials in mHealth. *Stat. Med.* **35** 1944–1971. [MR3513494](#) <https://doi.org/10.1002/sim.6847>
- MANCL, L. A. and DEROUEN, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics* **57** 126–134. [MR1833298](#) <https://doi.org/10.1111/j.0006-341X.2001.00126.x>
- MANCL, L. A. and LEROUX, B. G. (1996). Efficiency of regression estimates for clustered data. *Biometrics* **52** 500–511.
- PEARL, J. (2009). Causal inference in statistics: An overview. *Stat. Surv.* **3** 96–146. [MR2545291](#) <https://doi.org/10.1214/09-SS057>
- PEPE, M. S. and ANDERSON, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Comm. Statist. Simulation Comput.* **23** 939–951.
- 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](https://doi.org/10.1016/0270-0255(86)90088-6)
- ROSENBERGER, W. F., USCHNER, D. and WANG, Y. (2019). Randomization: The forgotten component of the randomized clinical trial. *Stat. Med.* **38** 1–12. [MR3887263](#) <https://doi.org/10.1002/sim.7901>
- RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. [MR0472152](#)
- SALEHEEN, N., ALI, A. A., HOSSAIN, S. M., SARKER, H., CHATTERJEE, S., MARLIN, B., ERTIN, E., AL'ABSI, M. and KUMAR, S. (2015). puffmarker: A multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15* 999–1010, ACM, New York, NY, USA. Available at <http://doi.acm.org/10.1145/2750858.2806897>.
- SARKER, H., TYBURSKI, M., RAHMAN, M. M., HOVSEPIAN, K., SHARMIN, M., EPSTEIN, D. H., PRESTON, K. L., FURR-HOLDEN, C. D., MILAM, A. et al. (2016). Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16* 4489–4501. ACM, Santa Clara, CA, USA.
- SARKER, H., HOVSEPIAN, K., CHATTERJEE, S., NAHUM-SHANI, I., MURPHY, S. A., SPRING, B., ERTIN, E., AL'ABSI, M., NAKAJIMA, M. et al. (2017). From markers to interventions: The case of just-in-time stress intervention. In *Mobile Health Sensors, Analytic Methods, and Applications* (J. M. Regh, S. A. Murphy and S. Kumar, eds.) Springer, Berlin.
- SPRUIJT-METZ, D. and NILSEN, W. (2014). Dynamic models of behavior for just-in-time adaptive interventions. *Pervasive Computing, IEEE* **13** 13–17.
- TCHETGEN TCHETGEN, E. J., GLYMOUR, M. M., WEUVE, J. and ROBINS, J. (2012). Specifying the correlation structure in inverse-probability-weighting estimation for repeated measures. *Epidemiology* **23** 644–646.
- VANDERWEELE, T. J., HONG, G., JONES, S. M. and BROWN, J. L. (2013). Mediation and spillover effects in group-randomized trials: A case study of the 4Rs educational intervention. *J. Amer. Statist. Assoc.* **108** 469–482. [MR3174634](#) <https://doi.org/10.1080/01621459.2013.779832>
- VANSTEELENDT, S. (2007). On confounding, prediction and efficiency in the analysis of longitudinal and cross-sectional clustered data. *Scand. J. Stat.* **34** 478–498. [MR2368794](#) <https://doi.org/10.1111/j.1467-9469.2006.00555.x>
- Wikipedia. F1 score—Wikipedia, the free encyclopedia, 2017. URL https://en.wikipedia.org/wiki/F1_score. [Online; accessed 23-May-2017].

ACCOUNTING FOR UNCERTAINTY ABOUT PAST VALUES IN PROBABILISTIC PROJECTIONS OF THE TOTAL FERTILITY RATE FOR MOST COUNTRIES

BY PEIRAN LIU* AND ADRIAN E. RAFTERY**

Department of Statistics, University of Washington, *prliu@uw.edu; **raftery@uw.edu

Since the 1940s, population projections have in most cases been produced using the deterministic cohort component method. However, in 2015, for the first time and in a major advance, the United Nations issued official probabilistic population projections for all countries based on Bayesian hierarchical models for total fertility and life expectancy. The estimates of these models and the resulting projections are conditional on the U.N.'s official estimates of past values. However, these past values are themselves uncertain, particularly for the majority of the world's countries that do not have long-standing high-quality vital registration systems, when they rely on surveys and censuses with their own biases and measurement errors. This paper extends the U.N. model for projecting future total fertility rates to take account of uncertainty about past values. This is done by adding an additional level to the hierarchical model to represent the multiple data sources, in each case estimating their bias and measurement error variance. We assess the method by out-of-sample predictive validation. While the prediction intervals produced by the extant method (which does not account for this source of uncertainty) have somewhat less than nominal coverage, we find that our proposed method achieves closer to nominal coverage. The prediction intervals become wider for countries for which the estimates of past total fertility rates rely heavily on surveys rather than on vital registration data, especially in high fertility countries.

REFERENCES

- ABEL, G. J., BARAKAT, B., SAMIR, K. C. and LUTZ, W. (2016). Meeting the sustainable development goals leads to lower world population growth. *Proc. Natl. Acad. Sci. USA* **113** 14294–14299.
- ALDERS, M., KEILMAN, N. and CRUIJSEN, H. (2007). Assumptions for long-term stochastic population forecasts in 18 European countries: Hypothèses de projections stochastiques à long terme des populations de 18 pays européens. *Eur. J. Popul.* **23** 33–69. <https://doi.org/10.1007/s10680-006-9104-4>
- ALHO, J. M., JENSEN, S. E. H. and LASSILA, J. (2008). *Uncertain Demographics and Fiscal Sustainability*. Cambridge Univ. Press, Cambridge, U.K.
- ALHO, J., ALDERS, M., CRUIJSEN, H., KEILMAN, N., NIKANDER, T. and PHAM, D. Q. (2006). New forecast: Population decline postponed in Europe. *Stat. J. U.N. Econ. Comm. Eur.* **23** 1–10.
- ALKEMA, L., RAFTERY, A. E., GERLAND, P., CLARK, S. J., PELLETIER, F., BUETTNER, T. and HEILIG, G. K. (2011). Probabilistic projections of the total fertility rate for all countries. *Demography* **48** 815–839.
- ALKEMA, L., RAFTERY, A. E., GERLAND, P., CLARK, S. J. and PELLETIER, F. (2012). Estimating trends in the total fertility rate with uncertainty using imperfect data: Examples from West Africa. *Demogr. Res.* **26** 331–362.
- BOHK-EWALD, C., LI, P. and MYRSKYLA, M. (2018). Forecast accuracy hardly improves with method complexity when completing cohort fertility. *Proc. Natl. Acad. Sci. USA* **115** 9187–9192.
- BOOTH, H., PENNEC, S. and HYNDMAN, R. J. (2009). Stochastic population forecasting using functional data methods: The case of France. In *Annual Meeting of the International Union for the Scientific Study of Population, Marrakech, Morocco*.
- BRASS, W. (1964). *Uses of Census or Survey Data for the Estimation of Vital Rates*. United Nations, New York.
- BRASS, W. (2015). *Demography of Tropical Africa*. Princeton Univ. Press, Princeton, N.J.

- CAI, Y. (2008). An assessment of China's fertility level using the variable-R method. *Demography* **45** 271–281.
- CANNAN, E. (1895). The probability of cessation of growth of population in England and Wales during the next century. *Econ. J.* **5** 506–515.
- EDIEV, D. (2013). Comparative importance of the fertility model, the total fertility, the mean age and the standard deviation of age at childbearing in population projections. Paper presented to the International Population Conference, Busan, Korea. <https://www.iussp.org/en/event/17/programme/paper/3054>.
- FOSDICK, B. K. and RAFTERY, A. E. (2014). Regional probabilistic fertility forecasting by modeling between-country correlations. *Demogr. Res.* **30** 1011–1034. <https://doi.org/10.4054/demres.2014.30.35>
- GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* 457–472.
- GERLAND, P., RAFTERY, A. E., ŠEVČÍKOVÁ, H., LI, N., GU, D., SPOORENBERG, T., ALKEMA, L., FOSDICK, B. K., CHUNN, J. L. et al. (2014). World population stabilization unlikely this century. *Science* **346** 234–237.
- GOODKIND, D. M. (2004). China's missing children: The 2000 census underreporting surprise. *Popul. Stud.* **58** 281–295.
- KEYFITZ, N. (1981). The limits of population forecasting. *Popul. Dev. Rev.* **7** 579–593.
- LEE, R. D. (1993). Modeling and forecasting the time series of US fertility: Age distribution, range, and ultimate level. *Int. J. Forecast.* **9** 187–202. [https://doi.org/10.1016/0169-2070\(93\)90004-7](https://doi.org/10.1016/0169-2070(93)90004-7)
- LEE, R. D. and BULATAO, R. A. (2000). *Beyond Six Billion: Forecasting the World's Population*. National Academies Press, Washington.
- LEE, R. D. and TULJAPURKAR, S. (1994). Stochastic population forecasts for the United States: Beyond high, medium, and low. *J. Amer. Statist. Assoc.* **89** 1175–1189.
- LIU, P. and RAFTERY, A. E. (2020). Supplement to “Accounting for uncertainty about past values in probabilistic projections of the total fertility rate for most countries.” <https://doi.org/10.1214/19-AOS1294SUPP>.
- LUTZ, W. and SAMIR, K. C. (2010). Dimensions of global population projections: What do we know about future population trends and structures? *Philos. Trans. R. Soc. B* **365** 2779–2791.
- MYRSKYLÄ, M., GOLDSTEIN, J. R. and CHENG, Y. A. (2013). New cohort fertility forecasts for the developed world: Rises, falls, and reversals. *Popul. Dev. Rev.* **39** 31–56.
- MERLI, M. G. and RAFTERY, A. E. (2000). Are births underreported in rural China? Manipulation of statistical records in response to China's population policies. *Demography* **37** 109–126.
- MURRAY, C. J., CALLENDER, C. S., KULIKOFF, X. R., SRINIVASAN, V., ABATE, D., ABATE, K. H., ABAY, S. M., ABBASI, N., ABBASTABAR, H. et al. (2018). Population and fertility by age and sex for 195 countries and territories, 1950–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392** 1995–2051.
- NATIONAL POPULATION COMMISSION, F. R. O. N. (2009). Nigeria demographic and health survey 2008.
- NEAL, R. M. (2003). Slice sampling. *Ann. Statist.* **31** 705–767. [MR1994729 https://doi.org/10.1214/aos/1056562461](https://doi.org/10.1214/aos/1056562461)
- PRESTON, S. H., HEUVELINE, P. and GUILLOT, M. (2000). *Demography: Measuring and Modeling Population Processes*. Blackwell, Malden, MA.
- PULLUM, T. W., SCHOUMAKER, B., BECKER, S. and BRADLEY, S. E. (2013). An assessment of DHS estimates of fertility and under-five mortality. In *International Population Conference of the International Union for the Scientific Study of Population (IUSSP), Session 132: Data Quality in Demographic Surveys, August 28*.
- RAFTERY, A. E., ALKEMA, L. and GERLAND, P. (2014). Bayesian population projections for the United Nations. *Statist. Sci.* **29** 58–68. [MR3201847 https://doi.org/10.1214/13-STS419](https://doi.org/10.1214/13-STS419)
- RAFTERY, A. E., LALIC, N. and GERLAND, P. (2014). Joint probabilistic projection of female and male life expectancy. *Demogr. Res.* **30** 795–822. <https://doi.org/10.4054/DemRes.2014.30.27>
- RAFTERY, A. E. and LEWIS, S. M. (1996). Implementing MCMC. In *Markov Chain Monte Carlo in Practice. Interdisciplinary Statistics*. CRC Press, London. [MR1397966 https://doi.org/10.1007/978-1-4899-4485-6](https://doi.org/10.1007/978-1-4899-4485-6)
- RAFTERY, A. E., GNEITING, T., BALABDAOUI, F. and POLAKOWSKI, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133** 1155–1174.
- RAFTERY, A. E., CHUNN, J. L., GERLAND, P. and ŠEVČÍKOVÁ, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography* **50** 777–801. <https://doi.org/10.1007/s13524-012-0193-x>
- RETFERFORD, R. D., CHOE, M. K., CHEN, J., XIRU, L. and HONGYAN, C. (2005). How far has fertility in China really declined? *Popul. Dev. Rev.* **31** 57–84.
- SCHMERTMANN, C., ZAGHENI, E., GOLDSTEIN, J. R. and MYRSKYLÄ, M. (2014). Bayesian forecasting of cohort fertility. *J. Amer. Statist. Assoc.* **109** 500–513. [MR3223728 https://doi.org/10.1080/01621459.2014.881738](https://doi.org/10.1080/01621459.2014.881738)
- SCHOUMAKER, B. (2010). Reconstructing fertility trends in sub-Saharan Africa by combining multiple surveys affected by data quality problems. In *Proceedings of the 2010 Annual Meeting of the Population Association of America*.

- SCHOUMAKER, B. (2011). Omissions of births in DHS birth histories in sub-Saharan Africa: Measurement and determinants. In *Proceedings of the 2011 Annual Meeting of the Population Association of America* **31**.
- SCHOUMAKER, B. (2014). Quality and consistency of DHS fertility estimates, 1990 to 2012. ICF International Rockville.
- ŠEVČÍKOVÁ, H., ALKEMA, L. and RAFTERY, A. E. (2011). bayesTFR: An R package for probabilistic projections of the total fertility rate. *J. Stat. Softw.* **43** 1–29.
- ŠEVČÍKOVÁ, H. and RAFTERY, A. E. (2016). bayesPop: Probabilistic population projections. *J. Stat. Softw.* **75** 1–29.
- ŠEVČÍKOVÁ, H., LI, N., KANTOROVÁ, V., GERLAND, P. and RAFTERY, A. E. (2016). Age-specific mortality and fertility rates for probabilistic population projections. In *Dynamic Demographic Analysis. Springer Ser. Demogr. Methods Popul. Anal.* **39** 285–310. Springer, Cham. [MR3525591](#) https://doi.org/10.1007/978-3-319-26603-9_15
- STOTO, M. A. (1983). The accuracy of population projections. *J. Amer. Statist. Assoc.* **78** 13–20. <https://doi.org/10.1080/01621459.1983.10477916>
- U. S. CENSUS BUREAU, P. D. (2017). Annual estimates of the resident population: April 1, 2010 to July 1, 2017.
- UNITED NATIONS (2008). World population prospects: The 2008 revision. United Nations, New York.
- UNITED NATIONS (2015a). World population prospects: The 2015 revision. United Nations, New York.
- UNITED NATIONS (2015b). World fertility data. <http://www.un.org/en/development/desa/population/publications/dataset/fertility/wfd2015.shtml>.
- UNITED NATIONS (2017). World population prospects: The 2017 revision. United Nations, New York.
- UNITED NATIONS (2019). The Sustainable Development Goals Report 2019. <https://unstats.un.org/sdgs/report/2019>.
- UNICEF (2016). UNICEF Annual Report 2015.
- WHELDON, M. C. (2019). Personal communication.
- WHELPTON, P. K. (1928). Population of the United States, 1925–1975. *Amer. J. Sociol.* **31** 253–270.
- WHELPTON, P. K. (1936). An empirical method for calculating future population. *J. Amer. Statist. Assoc.* **31** 457–473.
- YI, Z. (1996). Is fertility in China in 1991–92 far below replacement level? *Popul. Stud.* **50** 27–34.
- ZHAI, Z. and CHEN, W. (2007). Research of China's total fertility rates in late 1990s (in Chinese). *Demogr. Res.* **31** 19–31.

ESTIMATION OF DYADIC CHARACTERISTICS OF FAMILY NETWORKS USING SAMPLE SURVEY DATA

BY CHRIS SKINNER^{*} AND FIONA STEELE[†]

Department of Statistics, London School of Economics and Political Science, ^{}c.j.skinner@lse.ac.uk; [†]f.a.steele@lse.ac.uk*

We consider the use of sample survey data to estimate dyadic characteristics of family networks, with an application to noncoresident parent–child dyads. We suppose that survey respondents report either from a parent or child perspective about a dyad, depending on their membership of the dyad. We construct separate estimators of common dyadic characteristics using data from both a parent and a child perspective and show how comparisons of these estimators can shed light on data quality issues including differential missingness and reporting error. In our application we find that a simple missingness model explains some striking patterns of discrepancies between the estimators and consider the use of poststratification and a related redefinition of count variables to adjust for these discrepancies. We also develop approaches to combining the separate estimators efficiently to estimate means and frequency distributions within subpopulations.

REFERENCES

- ALTONJI, J. G., HAYASHI, F. and KOTLIKOFF, L. J. (1997). Parental altruism and inter vivos transfers: Theory and evidence. *J. Polit. Econ.* **105** 1121–1166.
- BRANDT, M., HABERKERN, K. and SZYDLIK, M. (2009). Intergenerational help and care in Europe. *Eur. Sociol. Rev.* **25** 585–601.
- CHAN, T. W. and ERMISCH, J. (2011). Intergenerational exchange of instrumental support: Dynamic evidence from the British Household Survey Working paper, Univ. Oxford.
- FRANK, O. (1977). Survey sampling in graphs. *J. Statist. Plann. Inference* **1** 235–264. [MR0494613 https://doi.org/10.1016/0378-3758\(77\)90011-8](https://doi.org/10.1016/0378-3758(77)90011-8)
- FULLER, W. A. (2009). *Sampling Statistics*. Wiley, Hoboken, NJ.
- HANDCOCK, M. S. and GILE, K. J. (2007). Modeling social networks with sampled or missing data. Working Paper 75, Centre for Statistics and the Social Sciences, Univ. Washington.
- HANDCOCK, M. S. and GILE, K. J. (2010). Modeling social networks from sampled data. *Ann. Appl. Stat.* **4** 5–25. [MR2758082 https://doi.org/10.1214/08-AOAS221](https://doi.org/10.1214/08-AOAS221)
- HATTON, T. J. (2005). Explaining trends in UK immigration. *J. Popul. Econ.* **18** 719–740.
- HENRETTA, J. C., GRUNDY, E. and HARRIS, S. (2002). The influence of socio-economic and health differences on parents’ provision of help to adult children: A British-United States comparison. *Ageing Soc.* **22** 441–458.
- ISER (2018). *British Household Panel Survey: Waves 1–18, 1991–2009*, 8th ed. Univ. Essex, Institute for Social and Economic Research [original data producer(s)], Colchester, Essex: UK Data Archive [distributor]. SN: 5151.
- LOHR, S. L. (2010). *Sampling: Design and Analysis*, 2nd ed. Brooks/Cole, Cengage Learning, Boston, MA. [MR3057878](https://doi.org/10.1016/B978-0-7607-6526-7)
- MANDEMAKERS, J. J. and DYKSTRA, P. A. (2008). Discrepancies in parent’s and adult child’s reports of support and contact. *J. Marriage Fam.* **70** 495–506.
- RENDALL, M. S. and BALL, D. J. (2004). Immigration, emigration and the ageing of the overseas-born population in the United Kingdom. *Popul. Trends* **116** 18–27.
- RENDALL, M. S., CLARKE, L., PETERS, E. H., RANJIT, N. and VERROPOULOU, G. (1999). Incomplete reporting of men’s fertility in the United States and Britain: A research note. *Demography* **36** 135–144.
- SHAPIRO, A. (2004). Revisiting the generation gap: Exploring the relationships of parent/adult-child dyads. *Int. J. Aging Hum. Dev.* **58** 127–146.
- SKINNER, C. and STEELE, F. (2020). Supplement to “Estimation of dyadic characteristics of family networks using sample survey data.” <https://doi.org/10.1214/19-AOAS1308SUPP>.

- SNIJDERS, T. A. B. and KENNY, D. A. (1999). The social relations model for family data: A multilevel approach. *Pers. Relatsh.* **6** 471–486.
- SORENSEN, E. (1997). A national profile of nonresident fathers and their ability to pay child support. *J. Marriage Fam.* **59** 785–808.
- STYKES, J. B., MANNING, W. D. and BROWN, S. L. (2013). Nonresident fathers and formal child support: Evidence from the CPS, the NSFG and the SIPP. *Demogr. Res.* **29** 1299–1330.
- TAYLOR, M. F., BRICE, J., BUCK, N. and PRENTICE-LANE, E. (2010). *British Household Panel Survey User Manual, Volume A: Introduction, Technical Report and Appendices*. Univ. Essex, Colchester.
- VALLIANT, R., DEVER, J. A. and KREUTER, F. (2013). *Practical Tools for Designing and Weighting Survey Samples. Statistics for Social and Behavioral Sciences*. Springer, New York. **MR3088726** <https://doi.org/10.1007/978-1-4614-6449-5>
- WIDMER, E. D., AEBY, G. and SAPIN, M. (2013). Collecting family network data. *Int. Rev. Sociol.* **23** 27–46.

SEQUENTIAL IMPORTANCE SAMPLING FOR MULTIREOLUTION KINGMAN-TAJIMA COALESCENT COUNTING

BY LORENZO CAPPELLO^{1,*} AND JULIA A. PALACIOS^{1,2,†}

¹Department of Statistics, Stanford University

²Department of Biomedical Data Science, Stanford University, *cappello@stanford.edu; †juliapr@stanford.edu

Statistical inference of evolutionary parameters from molecular sequence data relies on coalescent models to account for the shared genealogical ancestry of the samples. However, inferential algorithms do not scale to available data sets. A strategy to improve computational efficiency is to rely on simpler coalescent and mutation models, resulting in smaller hidden state spaces. An estimate of the cardinality of the state space of genealogical trees at different resolutions is essential to decide the best modeling strategy for a given dataset. To our knowledge, there is neither an exact nor approximate method to determine these cardinalities. We propose a sequential importance sampling algorithm to estimate the cardinality of the sample space of genealogical trees under different coalescent resolutions. Our sampling scheme proceeds sequentially across the set of combinatorial constraints imposed by the data which, in this work, are completely linked sequences of DNA at a non-recombining segment. We analyze the cardinality of different genealogical tree spaces on simulations to study the settings that favor coarser resolutions. We apply our method to estimate the cardinality of genealogical tree spaces from mtDNA data from the 1000 genomes and a sample from a Melanesian population at the β -globin locus.

REFERENCES

- 1000 GENOMES PROJECT CONSORTIUM (2015). A global reference for human genetic variation. *Nature* **526** 68 EP.
- ANDERSON, S., BANKIER, A. T., BARRELL, B. G., DE BRUIJN, M. H., COULSON, A. R., DROUIN, J., EPERON, I. C., NIERLICH, D. P., ROE, B. A. et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature* **290** 457.
- ANDREWS, R. M., KUBACKA, I., CHINNERY, P. F., LIGHTOWLERS, R. N., TURNBULL, D. M. and HOWELL, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23** 147. <https://doi.org/10.1038/13779>
- BEHAR, D. M., VAN OVEN, M., ROSSET, S., METSPALU, M., LOOGVÄLI, E.-L., SILVA, N. M., KIVISILD, T., TORRONI, A. and VILLEMS, R. (2012). A “Copernican” reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.* **90** 675–684.
- BLANCHET, J. and RUDOLY, D. (2009). Rare event simulation and counting problems. In *Rare Event Simulation Using Monte Carlo Methods* 171–192. Wiley, Chichester. MR2730766 <https://doi.org/10.1002/9780470745403.ch8>
- BLITZSTEIN, J. and DIACONIS, P. (2010). A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Math.* **6** 489–522. MR2809836 <https://doi.org/10.1080/15427951.2010.557277>
- CAYLEY, A. (1856). Note sur une formule pour la reversion des séries. *J. Reine Angew. Math.* **52** 276–284. MR1578984 <https://doi.org/10.1515/crll.1856.52.276>
- CHATTERJEE, S. and DIACONIS, P. (2018). The sample size required in importance sampling. *Ann. Appl. Probab.* **28** 1099–1135. MR3784496 <https://doi.org/10.1214/17-AAP1326>
- CHEN, Y. and CHEN, Y. (2018). An efficient sampling algorithm for network motif detection. *J. Comput. Graph. Statist.* **27** 503–515. MR3863753 <https://doi.org/10.1080/10618600.2017.1391696>
- CHEN, Y., DIACONIS, P., HOLMES, S. P. and LIU, J. S. (2005). Sequential Monte Carlo methods for statistical analysis of tables. *J. Amer. Statist. Assoc.* **100** 109–120. MR2156822 <https://doi.org/10.1198/016214504000001303>

- DIACONIS, P. (2018). Sequential importance sampling for estimating the number of perfect matchings in bipartite graphs: An ongoing conversation with Laci. Preprint.
- DISANTO, F. and WIEHE, T. (2013). Exact enumeration of cherries and pitchforks in ranked trees under the coalescent model. *Math. Biosci.* **242** 195–200. [MR3068684 https://doi.org/10.1016/j.mbs.2013.01.010](https://doi.org/10.1016/j.mbs.2013.01.010)
- DRUMMOND, A. J., SUCHARD, M. A., XIE, D. and RAMBAUT, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29** 1969–1973.
- FERRETTI, L., LEDDA, A., WIEHE, T., ACHAZ, G. and RAMOS-ONSINS, S. E. (2017). Decomposing the site frequency spectrum: The impact of tree topology on neutrality tests. *Genetics* **207** 229–240. <https://doi.org/10.1534/genetics.116.188763>
- FULLERTON, S. M., HARDING, R. M., BOYCE, A. J. and CLEGG, J. B. (1994). Molecular and population genetic analysis of allelic sequence diversity at the human beta-globin locus. *Proc. Natl. Acad. Sci. USA* **91** 1805–1809.
- GAO, F. and KEINAN, A. (2016). Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. *Genetics* **202** 235–245.
- GATTEPAILLE, L., GÜNTHER, T. and JAKOBSSON, M. (2016). Inferring past effective population size from distributions of coalescent times. *Genetics* **204** 1191–1206. <https://doi.org/10.1534/genetics.115.185058>
- GRIFFITHS, R. C. (1987). Counting genealogical trees. *J. Math. Biol.* **25** 423–431. [MR0908383 https://doi.org/10.1007/BF00277166](https://doi.org/10.1007/BF00277166)
- GRIFFITHS, R. C. and MARJORAM, P. (1997). An ancestral recombination graph. In *Progress in Population Genetics and Human Evolution* (Minneapolis, MN, 1994). *IMA Vol. Math. Appl.* **87** 257–270. Springer, New York. [MR1493031 https://doi.org/10.1007/978-1-4757-2609-1_16](https://doi.org/10.1007/978-1-4757-2609-1_16)
- GRIFFITHS, R. C. and TAVARÉ, S. (1999). The ages of mutations in gene trees. *Ann. Appl. Probab.* **9** 567–590. [MR1722273 https://doi.org/10.1214/aoap/1029962804](https://doi.org/10.1214/aoap/1029962804)
- GUSFIELD, D. (1991). Efficient algorithms for inferring evolutionary trees. *Networks* **21** 19–28. [MR1083125 https://doi.org/10.1002/net.3230210104](https://doi.org/10.1002/net.3230210104)
- GUSFIELD, D. (2014). *ReCombinatorics: The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks*. MIT Press, Cambridge, MA. With contributions from Charles H. Langley, Yun S. Song and Yufeng Wu. [MR3237544](https://doi.org/10.1017/9780521876223)
- HAMMERSLEY, J. M. and HANDSCOMB, D. C. (1965). *Monte Carlo Methods*. Methuen & Co., Ltd., London; Barnes & Noble, Inc., New York. [MR0223065](https://doi.org/10.1017/9780521876223)
- HARDING, R. M., FULLERTON, S. M., GRIFFITHS, R. C., BOND, J., COX, M. J., SCHNEIDER, J. A., MOULIN, D. S. and CLEGG, J. B. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60** 772–789.
- JERRUM, M. and SINCLAIR, A. (1996). The Markov chain Monte Carlo method: An approach to approximate counting and integration. In *Approximation Algorithms for NP-Hard Problems* 482–520.
- JERRUM, M. R., VALIANT, L. G. and VAZIRANI, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.* **43** 169–188. [MR0855970 https://doi.org/10.1016/0304-3975\(86\)90174-X](https://doi.org/10.1016/0304-3975(86)90174-X)
- KIMURA, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61** 893.
- KINGMAN, J. F. C. (1982). On the genealogy of large populations. *J. Appl. Probab. Special Vol.* **19A** 27–43. Essays in statistical science. [MR0633178](https://doi.org/10.1017/9780521876223)
- KNUTH, D. E. (1976). Mathematics and computer science: Coping with finiteness. *Science* **194** 1235–1242. [MR0534161 https://doi.org/10.1126/science.194.4271.1235](https://doi.org/10.1126/science.194.4271.1235)
- KNUTH, D. E. (2018). *Art of Computer Programming, Volume 4B, Fascicle 5: The: Mathematical Preliminaries Redux; Backtracking; Dancing Links*. Addison Wesley Professional.
- LIU, D., SHI, W., SHI, Y., WANG, D., XIAO, H., LI, W., BI, Y., WU, Y., LI, X. et al. (2013). Origin and diversity of novel avian influenza A H7N9 viruses causing human infection: Phylogenetic, structural, and coalescent analyses. *Lancet* **381** 1926–1932.
- MALLET, O., GASCUEL, F. and LAMBERT, A. (2018). Ranked tree shapes, nonrandom extinctions, and the loss of phylogenetic diversity. *Syst. Biol.* **67** 1025–1040. <https://doi.org/10.1093/sysbio/syy030>
- NORDBORG, M. (1998). On the probability of Neanderthal ancestry. *Am. J. Hum. Genet.* **63** 1237–1240. <https://doi.org/10.1086/302052>
- OWEN, A. B. (2013). Monte Carlo theory, methods and examples. Online.
- PALACIOS, J. A. and MININ, V. N. (2013). Gaussian process-based Bayesian nonparametric inference of population size trajectories from gene genealogies. *Biometrics* **69** 8–18. [MR3058047 https://doi.org/10.1111/biom.12003](https://doi.org/10.1111/biom.12003)
- PALACIOS, J. A., VÉBER, A., CAPPELLO, L., WANG, Z., WAKELEY, J. and RAMACHANDRAN, S. (2019). Bayesian estimation of population size changes by sampling Tajima’s trees. *Genetics* **213** 967–986.

- PARADIS, E., CLAUDE, J. and STRIMMER, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20** 289–290.
- ROSENBERG, N. A. and NORDBOG, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* **3** 380–390. <https://doi.org/10.1038/nrg795>
- SAINUDIIN, R., STADLER, T. and VÉBER, A. (2015). Finding the best resolution for the Kingman–Tajima coalescent: Theory and applications. *J. Math. Biol.* **70** 1207–1247. [MR3323594 https://doi.org/10.1007/s00285-014-0796-5](https://doi.org/10.1007/s00285-014-0796-5)
- SAINUDIIN, R. and VÉBER, A. (2018). Full likelihood inference from the site frequency spectrum based on the optimal tree resolution. *Theor. Popul. Biol.* **124** 1–40.
- SINCLAIR, A. (2012). *Algorithms for Random Generation and Counting: A Markov Chain Approach*. Springer Science & Business Media.
- SONG, S., PURSELL, Z. F., COPELAND, W. C., LONGLEY, M. J., KUNKEL, T. A. and MATHEWS, C. K. (2005). DNA precursor asymmetries in mammalian tissue mitochondria and possible contribution to mutagenesis through reduced replication fidelity. *Proc. Natl. Acad. Sci. USA* **102** 4990–4995.
- STEEL, M. (2016). *Phylogeny—Discrete and Random Processes in Evolution. CBMS-NSF Regional Conference Series in Applied Mathematics* **89**. SIAM, Philadelphia, PA. [MR3601108 https://doi.org/10.1137/1.9781611974485.ch1](https://doi.org/10.1137/1.9781611974485.ch1)
- TAJIMA, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105** 437–460.
- TAVARÉ, S. (2004). *Ancestral Inference in Population Genetics. Lectures on Probability Theory and Statistics: Ecole D'Eté de Probabilités de Saint-Flour XXXI-2001*. Springer.
- TERHORST, J., KAMM, J. A. and SONG, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49** 303–309.
- WANG, L., BOUCHARD-CÔTÉ, A. and DOUCET, A. (2015). Bayesian phylogenetic inference using a combinatorial sequential Monte Carlo method. *J. Amer. Statist. Assoc.* **110** 1362–1374. [MR3449032 https://doi.org/10.1080/01621459.2015.1054487](https://doi.org/10.1080/01621459.2015.1054487)
- WATTERSON, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7** 256–276. [MR0366430 https://doi.org/10.1016/0040-5809\(75\)90020-9](https://doi.org/10.1016/0040-5809(75)90020-9)

MULTIVIEW CLUSTER AGGREGATION AND SPLITTING, WITH AN APPLICATION TO MULTIOMIC BREAST CANCER DATA

BY ANTOINE GODICHON-BAGGIONI¹, CATHY MAUGIS-RABUSSEAU² AND ANDREA RAU³

¹*Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, antoine.godichon_baggioni@upmc.fr*

²*Institut de Mathématiques de Toulouse, UMR5219, Université de Toulouse, CNRS, INSA Toulouse, cathy.maugis@insa-toulouse.fr*

³*Université Paris-Saclay, INRAE, AgroParisTech, GABI, andrea.rau@inrae.fr*

Multiview data, which represent distinct but related groupings of variables, can be useful for identifying relevant and robust clustering structures among observations. A large number of multiview classification algorithms have been proposed in the fields of computer science and genomics; here, we instead focus on the task of merging or splitting an existing hard or soft cluster partition based on multiview data. This article is specifically motivated by an application involving multiomic breast cancer data from The Cancer Genome Atlas, where multiple molecular profiles (gene expression, microRNA expression, methylation and copy number alterations) are used to further subdivide the five currently accepted intrinsic tumor subtypes into distinct subgroups of patients. In addition, we investigate the performance of the proposed multiview splitting and aggregation algorithms, as compared to single- and concatenated-view alternatives, in a set of simulations. The multiview splitting and aggregation algorithms developed here are implemented in the *maskmeans* R package.

REFERENCES

- ACAR, E. and YENER, B. (2008). Unsupervised multiway data analysis: A literature survey. *IEEE Trans. Knowl. Data Eng.* **21** 6–20.
- BAUDRY, J.-P., MAUGIS, C. and MICHEL, B. (2012). Slope heuristics: Overview and implementation. *Stat. Comput.* **22** 455–470. [MR2865029 https://doi.org/10.1007/s11222-011-9236-1](https://doi.org/10.1007/s11222-011-9236-1)
- BAUDRY, J.-P., RAFTERY, A. E., CELEUX, G., LO, K. and GOTTARDO, R. (2010). Combining mixture components for clustering. *J. Comput. Graph. Statist.* **19** 332–353. [MR2758307 https://doi.org/10.1198/jcgs.2010.08111](https://doi.org/10.1198/jcgs.2010.08111)
- BEZDEK, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York. [MR0631231](https://doi.org/10.1007/978-1-4613-1231-1)
- BRAY, F., FERLAY, J., SOERJOMATARAM, I., SIEGEL, R. L., TORRE, L. A. and JEMAL, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.*
- CAI, X., NIE, F. and HUANG, H. (2013). Multi-view k-means clustering on big data. In *Twenty-Third International Joint Conference on Artificial Intelligence*.
- CHAO, G., SUN, S. and BI, J. (2017). A survey on multi-view clustering. Preprint. Available at [arXiv:1712.06246](https://arxiv.org/abs/1712.06246).
- CHEN, X., XU, J. Z., HUANG, X. and YE, Y. (2013). TW-k-means: Automated two-level variable weighting clustering algorithm for multiview data. *IEEE Trans. Knowl. Data Eng.* **25**.
- CIRIELLO, G., SINHA, R., HOADLEY, K. A., JACOBSEN, A. S., REVA, B., PEROU, C. M., SANDER, C. and SCHULTZ, N. (2013). The molecular diversity of Luminal A breast tumors. *Breast Cancer Research and Treatment* **131** 409–420.
- DUECK, D., MORRIS, Q. D. and FREY, B. J. (2005). Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics* **21** i144–i151.
- FENG, Q., JIANG, M., HANNIG, J. and MARRON, J. S. (2018). Angle-based joint and individual variation explained. *J. Multivariate Anal.* **166** 241–265. [MR3799646 https://doi.org/10.1016/j.jmva.2018.03.008](https://doi.org/10.1016/j.jmva.2018.03.008)

- FERRARO, M. B. and GIORDANI, P. (2015). A toolbox for fuzzy clustering using the R programming language. *Fuzzy Sets and Systems* **279** 1–16. [MR3392321 https://doi.org/10.1016/j.fss.2015.05.001](https://doi.org/10.1016/j.fss.2015.05.001)
- FLEISCHER, T., KLAJIC, J., AURE, M. R., LOUHIMO, R., PLADSEN, A. V., OTTESTAD, L., TOULEIMAT, N., LAAKSO, M., HALVORSEN, A. R. et al. (2017). DNA methylation signature (SAM40) identifies subgroups of the Luminal A breast cancer samples with distinct survival. *Oncotarget* **8** 1074–1082.
- FRATELLO, M., CAIAZZO, G., TROJSI, F., RUSSO, A., TEDESCHI, G., TAGLIAFERRI, R. and ESPOSITO, F. (2017). Multi-view ensemble classification of brain connectivity images for neurodegeneration type discrimination. *Neuroinformatics* **15** 199–213.
- GAYNANOVA, I. and LI, G. (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics* **75** 1121–1132.
- GODICHON-BAGGIONI, A., MAUGIS-RABUSSEAU, C. and RAU, A. (2020). Supplement to “Multiview cluster aggregation and splitting, with an application to multiomic breast cancer data.” <https://doi.org/10.1214/19-AOAS1317SUPP>.
- GU, Z., EILS, R. and SCHLESNER, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32** 2847–2849.
- HAMID, J. S., HI, P., ROSLIN, N. M., LING, V., GREENWOOD, C. M. T. and BEYENE, J. (2009). Data integration in genetics and genomics: Methods and challenges. *Human Genomics* 869093.
- HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* 193–218.
- KAILING, K., KRIEGLER, H.-P., PRYAKHIN, A. and SCHUBERT, M. (2004). Clustering multi-represented objects with noise. In *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-04)* 394–403.
- KANEHISA, M., SATO, Y., KAWASHIMA, M., FURUMICHI, M. and TANABE, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44** D457–D462.
- KOUTSONIKOLA, V. A. and VAKALI, A. I. (2009). A fuzzy bi-clustering approach to correlate web users and pages. *International Journal of Knowledge and Web Intelligence* **1** 3–23.
- KUMAR, A. and DAUMÉ, H. (2011). A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* 393–400.
- KUMAR, A., RAI, P. and DAUME, H. (2011). Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems* 1413–1421.
- LIU, X., JI, S., GLÄNZEL, W. and DE MOOR, B. (2012). Multiview partitioning via tensor methods. *IEEE Trans. Knowl. Data Eng.* **25** 1056–1069.
- LIU, J., LICHTENBERG, T., HOADLEY, K. A., POISSON, L. M., LAZAR, A. J., CHERNIACK, A. D., KOVATICH, A. J., BENZ, C. C., LEVINE, D. A. et al. (2018). An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173** 400–416.
- LOCK, E. F., HOADLEY, K. A., MARRON, J. S. and NOBEL, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7** 523–542. [MR3086429 https://doi.org/10.1214/12-AOAS597](https://doi.org/10.1214/12-AOAS597)
- MO, Q., WANG, S., SESHAN, V. E., OLSHEN, A. B., SCHULTZ, N., SANDER, C., POWERS, R. S., LADANYI, M. and SHEN, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. USA* **110** 4245–4250.
- PAQUET, E. R. and HALLETT, M. T. (2000). Absolute assignment of breast cancer intrinsic molecular subtype. *J. Natl. Cancer Inst.* **107**.
- PARKER, J. S., MULLINS, M., CHEANG, M. C. U., LEUNG, S., VODUC, D., VICKERY, T., DAVIES, S., FAURON, C., HE, X. et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27** 1160–1167.
- PENSA, R. G., ROBARDET, C. and BOULICAUT, J.-F. (2005). A bi-clustering framework for categorical data. In *European Conference on Principles of Data Mining and Knowledge Discovery* 643–650. Springer, Berlin.
- 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** 747–752.
- RAPPOPORT, N. and SHAMIR, R. (2018). Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic Acids Res.* **gky889**.
- RAU, A., FLISTER, M., RUI, H. and AUER, P. L. (2018). Exploring drivers of gene expression in the Cancer Genome Atlas. *Bioinformatics* **bty551**.
- SERRA, A., FRATELLO, M., FORTINO, V., RAICONI, G., TAGLIAFERRI, R. and GRECO, D. (2015). Mvda: A multi-view genomic data integration methodology. *BMC Bioinform.* **16** 261.
- SHEN, R., OLSHEN, A. B. and LADANYI, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25** 2906–2912.
- SHEN, R., MO, Q., SCHULTZ, N., SESHAN, V. E., OLSHEN, A. B., HUSE, J., LADANYI, M. and SANDER, C. (2012). Integrative subtype discovery in glioblastoma using icluster. *PLoS ONE* **7** e35236.

- TASKESSEN, E., HUISMAN, S. M. H., MAHFOUZ, A., KRIJTHE, J. H., DE RIDDER, J., VAN DE STOLPE, A., VAN DEN AKKER, E., VERHEAGH, W. and REINDERS, M. J. T. (2016). Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. *Sci. Rep.* **6**.
- THE CANCER GENOME ATLAS NETWORK (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70.
- THE CANCER GENOME ATLAS NETWORK, WEINSTEIN, J. N., COLLISSON, E. A., MILLS, G. B., MILLS SHAW, K. R., OZENBERGER, B. A., ELLROTT, K., SHMULEVICH, I., SANDER, C. et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45** 1113–1120.
- WANG, Y. and CHEN, L. (2017). Multi-view fuzzy clustering with minimax optimization for effective clustering of data from multiple sources. *Expert Syst. Appl.* **72** 457–466.
- WANG, B., MEZLINI, A. M., DEMIR, F., FIUME, M., TU, Z., BRUDNO, M., HAIBE-KAINS, B. and GOLDENBERG, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11** 333–337.
- WICKHAM, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.
- XU, C., TAO, D. and XU, C. (2013). A survey on multi-view learning. Preprint. Available at [arXiv:1304.5634](https://arxiv.org/abs/1304.5634).
- YANG, Z. and MICHAILIDIS, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* **32** 1–8.
- YANG, Y. and WANG, H. (2018). Multi-view clustering: A survey. *Big Data Mining and Analytics* **1** 83–107.
- YU, Q., RISK, B. B., ZHANG, K. and MARRON, J. S. (2017). Jive integration of imaging and behavioral data. *NeuroImage* **152** 38–49.
- ZAPPIA, L. and OSHLACK, A. (2018). Clustering trees: A visualisation for evaluating clusterings at multiple resolutions. *GigaScience* **7**.

SEMIPARAMETRIC BAYESIAN MARKOV ANALYSIS OF PERSONALIZED BENEFIT-RISK ASSESSMENT

BY DONGYAN YAN¹, SUBHARUP GUHA², CHUL AHN^{3,*} AND RAM TIWARI^{3,†}

¹Discovery & Development Statistics, Eli Lilly and Company, dyyr2@mail.missouri.edu

²Department of Biostatistics, University of Florida, s.guha@ufl.edu

³Division of Biostatistics, Center for Devices and Radiological Health, Office Surveillance and Biometrics, U.S. Food and Drug Administration, *chul.ahn@fda.hhs.gov; †ram.tiwari@fda.hhs.gov

The development of systematic and structured approaches to assess benefit–risk of medical products is a major challenge for regulatory decision makers. Existing benefit–risk methods depend only on the frequencies of mutually exclusive and exhaustive categories in which the subjects fall, and the responses of individuals are allowed to belong to any of the other categories during their postwithdrawal visits. In this article we introduce a semiparametric Bayesian Markov model (SBMM) that treats the withdrawal category as an absorbing state and analyzes subject-level data for multiple visits, accounting for any within-patient dependencies in the response profiles. A log-odds ratio model is used to model the subject-level effects by assuming a ratio of transition probabilities with respect to a “reference” category. A Dirichlet process is used as a semiparametric model for the subject-level effects to flexibly capture the underlying distributions of the personalized response profiles without making strong parametric assumptions. This also allows the borrowing of strength between the patients and achieves dimension reduction by allocating similar response profiles patterns into an unknown number of latent clusters. We analyze a motivating clinical trial dataset to assess the personalized benefit–risks in each arm and evaluate the aggregated benefits and risks associated with the drug Exalgo.

REFERENCES

- BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355. [MR0362614](#)
- BROWN, E. R. and IBRAHIM, J. G. (2003). Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59** 686–693. [MR2004274](#) <https://doi.org/10.1111/1541-0420.00079>
- CAMERLENGHI, F., LIJOI, A., ORBANZ, P. and PRÜNSTER, I. (2019). Distribution theory for hierarchical processes. *Ann. Statist.* **47** 67–92. [MR3909927](#) <https://doi.org/10.1214/17-AOS1678>
- CHEN, M.-H., SHAO, Q.-M. and IBRAHIM, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer Series in Statistics. Springer, New York. [MR1742311](#) <https://doi.org/10.1007/978-1-4612-1276-8>
- CHIB, S. and GREENBERG, E. (1994). Bayes inference in regression models with ARMA(p, q) errors. *J. Econometrics* **64** 183–206. [MR1310523](#) [https://doi.org/10.1016/0304-4076\(94\)90063-9](https://doi.org/10.1016/0304-4076(94)90063-9)
- CHIB, S., GREENBERG, E. and WINKELMANN, R. (1998). Posterior simulation and Bayes factors in panel count data models. *J. Econometrics* **86** 33–54.
- CHUANG-STEIN, C., MOHBERG, N. R. and SINKULA, M. S. (1991). Three measures for simultaneously evaluating benefits and risks using categorical data from clinical trials. *Stat. Med.* **10** 1349–1359. <https://doi.org/10.1002/sim.4780100904>
- CLAIBORNE, A. B., ENGLISH, R. A. and CARUSO, D. (2014). *Characterizing and Communicating Uncertainty in the Assessment of Benefits and Risks of Pharmaceutical Products: Workshop Summary*. National Academies Press, Washington, DC.
- COSTA, M. J. and DRURY, T. (2018). Bayesian joint modelling of benefit and risk in drug development. *Pharm. Stat.* **17** 248–263.

Key words and phrases. Clinical trials, Dirichlet process, Exalgo, Gibbs sampling, log-odds ratio model, Metropolis–Hastings, SBMM.

- COSTA, M. J., HE, W., JEMIAI, Y., ZHAO, Y. and DI CASOLI, C. (2017). The case for a Bayesian approach to benefit–risk assessment: Overview and future directions. *Ther. Innov. Regul. Sci.* **51** 568–574.
- CUI, S., ZHAO, Y. and TIWARI, R. C. (2016). Bayesian approach to personalized benefit–risk assessment. *Stat. Biopharm. Res.* **8** 316–324.
- DAHL, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics* 201–218. Cambridge Univ. Press, Cambridge.
- ENTSUAH, R. and GORMAN, J. M. (2002). Global benefit–risk assessment of antidepressants: Venlafaxine XR and fluoxetine. *J. Psychiatr. Res.* **36** 111–118.
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. [MR0350949](#)
- GELBER, R. D., GELMAN, R. S. and GOLDBIRSH, A. (1989). A quality-of-life-oriented endpoint for comparing therapies. *Biometrics* **45** 781–795.
- GHOSAL, S., GHOSH, J. K. and RAMAMOORTHY, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27** 143–158. [MR1701105](#) <https://doi.org/10.1214/aos/1018031105>
- GHOSH, P., BASU, S. and TIWARI, R. C. (2009). Bayesian analysis of cancer rates from SEER program using parametric and semiparametric jointpoint regression models. *J. Amer. Statist. Assoc.* **104** 439–452. [MR2751429](#) <https://doi.org/10.1198/jasa.2009.0038>
- GLASZIOU, P. P., SIMES, R. J. and GELBER, R. D. (1990). Quality adjusted survival analysis. *Stat. Med.* **9** 1259–1276.
- GØTZSCHE, P. C. and JØRGENSEN, A. W. (2011). Opening up data at the European Medicines Agency. *BMJ, Br. Med. J.* **342**.
- GUHA, S. (2008). Posterior simulation in the generalized linear mixed model with semiparametric random effects. *J. Comput. Graph. Statist.* **17** 410–425. [MR2439966](#) <https://doi.org/10.1198/106186008X319854>
- HO, M.-W., TU, W., GHOSH, P. and TIWARI, R. C. (2013). A nested Dirichlet process analysis of cluster randomized trial data with application in geriatric care assessment. *J. Amer. Statist. Assoc.* **108** 48–68. [MR3174602](#) <https://doi.org/10.1080/01621459.2012.734164>
- HO, M., SAHA, A., MCCLEARY, K. K., LEVITAN, B., CHRISTOPHER, S., ZANDLO, K., BRAITHWAITE, R. S. and HAUBER, A. B. (2016). A framework for incorporating patient preferences regarding benefits and risks into regulatory assessment of medical technologies. *Value Health* **19** 746–750.
- HOLDEN, W. L., JUHAERI, J. and DAI, W. (2003). Benefit–risk analysis: A proposal using quantitative methods. *Pharmacoepidemiol. Drug Saf.* **12** 611–616.
- HUGHES, D., WADDINGHAM, E., MT ISA, S., GOGINSKY, A., CHAN, E., DOWNEY, G. F., HALL-GREEN, C. E., HOCKLEY, K. S., JUHAERI, J. et al. (2016). Recommendations for benefit–risk assessment methodologies and visual representations. *Pharmacoepidemiol. Drug Saf.* **25** 251–262.
- IBRAHIM, J. G. and CHEN, M.-H. (2000). Power prior distributions for regression models. *Statist. Sci.* **15** 46–60. [MR1842236](#) <https://doi.org/10.1214/ss/1009212673>
- ISHWARAN, H. and ZAREPOUR, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statist. Sinica* **12** 941–963. [MR1929973](#)
- NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. [MR1823804](#) <https://doi.org/10.2307/1390653>
- NORTON, J. D. (2011). A longitudinal model and graphic for benefit–risk analysis, with case study. *Drug Inf. J.* **45** 741–747.
- PAYNE, J. T. and LOKEN, M. K. (1975). A survey of the benefits and risks in the practice of radiology. *CRC Crit. Rev. Clin. Radiol. Nucl. Med.* **6** 425–439.
- PRITCHETT, Y. and TAMURA, R. (2008). The application of global benefit–risk assessment in clinical trial design and some statistical considerations. *Pharm. Stat.* **7** 170–178.
- RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2008). The nested Dirichlet process. *J. Amer. Statist. Assoc.* **103** 1131–1144. [MR2528831](#) <https://doi.org/10.1198/016214508000000553>
- SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. [MR1309433](#)
- SETHURAMAN, J. and TIWARI, R. C. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Statistical Decision Theory and Related Topics, III, Vol. 2 (West Lafayette, Ind., 1981)* 305–315. Academic Press, New York. [MR0705321](#)
- SHAHBABA, B. and NEAL, R. (2009). Nonlinear models using Dirichlet process mixtures. *J. Mach. Learn. Res.* **10** 1829–1850. [MR2540778](#)
- YAN, D., GUHA, S., AHN, C. and TIWARI, R. C. (2020). Supplement to “Semiparametric Bayesian Markov analysis of personalized benefit–risk assessment.” <https://doi.org/10.1214/20-AOAS1323SUPP>.
- ZEGER, S. L. and KARIM, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86. [MR1137101](#)
- ZHAO, Y., ZALKIKAR, J., TIWARI, R. C. and LAVANGE, L. M. (2014). A Bayesian approach for benefit–risk assessment. *Stat. Biopharm. Res.* **6** 326–337.

ESTIMATION AND INFERENCE IN METABOLOMICS WITH NONRANDOM MISSING DATA AND LATENT FACTORS

BY CHRIS MCKENNAN¹, CAROLE OBER² AND DAN NICOLAE³

¹*Department of Statistics, University of Pittsburgh, chm195@pitt.edu*

²*Department of Human Genetics, University of Chicago, c-ober@bsd.uchicago.edu*

³*Department of Statistics, University of Chicago, nicolae@galton.uchicago.edu*

High-throughput metabolomics data are fraught with both nonignorable missing observations and unobserved factors that influence a metabolite's measured concentration, and it is well known that ignoring either of these complications can compromise estimators. However, current methods to analyze these data can only account for the missing data or unobserved factors, but not both. We therefore developed MetabMiss, a statistically rigorous method to account for both nonrandom missing data and latent factors in high-throughput metabolomics data. Our methodology does not require the practitioner specify a likelihood for the missing data, and makes investigating the relationship between the metabolome and tens, or even hundreds, of phenotypes computationally tractable. We demonstrate the fidelity of MetabMiss's estimates using both simulated and real metabolomics data and prove their asymptotic correctness when the sample size and number of metabolites grows to infinity.

REFERENCES

- AFSHIN, A., SUR, P. J., FAY, K. A., CORNABY, L., FERRARA, G., SALAMA, J. S., MULLANY, E. C., ABATE, K. H., ABBAFATI, C. et al. (2019). Health effects of dietary risks in 195 countries, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **393** 1958–1972.
- BALATSINO, L., GIOACCHINO, G. D., SABATINO, G., CAVALLUCCI, E., CARUSO, R., GABRIELE, E., RAMONDO, S., GIAMPAOLO, L. D., VERNA, N. et al. (2004). Asthma worsened by benzoate contained in some antiasthmatic drugs. *International Journal of Immunopathology and Pharmacology* **17** 225–226.
- BAUM, C., SCHAFFER, M. and STILLMAN, S. (2003). Instrumental variables and GMM: Estimation and testing. *Stata Journal* **3** 1–31.
- BISGAARD, H., VISSING, N. H., CARSON, C. G., BISCHOFF, A. L., FØLSGAARD, N. V., KREINERMØLLER, E., CHAWES, B. L. K., STOKHOLM, J., PEDERSEN, L. et al. (2013). Deep phenotyping of the unselected COPSAC2010 birth cohort study. *Clinical and Experimental Allergy* **43** 1384–1394.
- BOUHIFD, M., BEGER, R., FLYNN, T., GUO, L., HARRIS, G., HOGBERG, H., KADDURAH-DAOUK, R., KAMP, H., KLEENSANG, A. et al. (2015). Quality assurance of metabolomics. *ALTEX* **32** 319–326.
- BROADHURST, D., GOODACRE, R., REINKE, S. N., KULIGOWSKI, J., WILSON, I. D., LEWIS, M. R. and DUNN, W. B. (2018). Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics* **14** 72. <https://doi.org/10.1007/s11306-018-1367-3>
- BROWN, B. W. and NEWAY, W. K. (2002). Generalized method of moments, efficient bootstrapping, and improved inference *Journal of Business & Economic Statistics* **20** 507–517. [MR1945606 https://doi.org/10.1198/073500102288618649](https://doi.org/10.1198/073500102288618649)
- BUJA, A. and EYUBOGLU, N. (1992). Remarks on parallel analysis. *Multivar. Behav. Res.* **27** 509–540.
- CHEN, L. S., WANG, J., WANG, X. and WANG, P. (2017). A mixed-effects model for incomplete data from labeling-based quantitative proteomics experiments. *The Annals of Applied Statistics* **11** 114–138. [MR3634317 https://doi.org/10.1214/16-AOAS994](https://doi.org/10.1214/16-AOAS994)
- CHOI, J. Y., LEE, H. Y., HUR, J., KIM, K. H., KANG, J. Y., RHEE, C. K. and LEE, S. Y. (2018). TRPV1 blocking alleviates airway inflammation and remodeling in a chronic asthma murine model. *Allergy, Asthma & Immunology Research* **10** 216–224.

- DAVIDSON, R. and MACKINNON, J. G. (2003). *Econometric Theory and Methods*. Oxford Univ. Press, London.
- DE LIVERA, A. M., DIAS, D. A., DE SOUZA, D., RUPASINGHE, T., PYKE, J., TULL, D., ROESSNER, U., MCCONVILLE, M. and SPEED, T. P. (2012). Normalizing and integrating metabolomics data. *Analytical Chemistry* **84** 10768–10776.
- DE LIVERA, A. M., SYSI-AHO, M., JACOB, L., GAGNON-BARTSCH, J. A., CASTILLO, S., SIMPSON, J. A. and SPEED, T. P. (2015). Statistical methods for handling unwanted variation in metabolomics data. *Analytical Chemistry* **87** 3606–3615.
- DO, K. T., WAHL, S., RAFFLER, J., MOLNOS, S., LAIMIGHOFER, M., ADAMSKI, J., SUHRE, K., STRAUCH, K., PETERS, A. et al. (2018). Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* **14** 128–128.
- DUBUIS, S., ORTMAYR, K. and ZAMPIERI, M. (2018). A framework for large-scale metabolome drug profiling links coenzyme A metabolism to the toxicity of anti-cancer drug dichloroacetate. *Communications Biology* **1** 101. <https://doi.org/10.1038/s42003-018-0111-x>
- FINKELSTEIN, J. L., PRESSMAN, E. K., COOPER, E. M., KENT, T. R., BAR, H. Y. and O'BRIEN, K. O. (2015). Vitamin D status affects serum metabolomic profiles in pregnant adolescents. *Reproductive Sciences* **22** 685–695.
- FRANKS, A. M., AIROLDI, E. M. and RUBIN, D. B. (2016). Non-standard conditionally specified models for non-ignorable missing data.
- FRIAS, B. and MERIGHI, A. (2016). Capsaicin, nociception and pain. *Molecules (Basel, Switzerland)* **21** 797.
- GAGNON-BARTSCH, J. A., JACOB, L. and SPEED, T. P. (2013). Removing unwanted variation from high dimensional data with negative controls. Technical Report, UC Berkeley.
- GAGNON-BARTSCH, J. A. and SPEED, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13** 539–552.
- HALL, P. and HOROWITZ, J. L. (1996). Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica* **64** 891–916. [MR1399222 https://doi.org/10.2307/2171849](https://doi.org/10.2307/2171849)
- HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. [MR0666123 https://doi.org/10.2307/1912775](https://doi.org/10.2307/1912775)
- HANSEN, B. E. and WEST, K. D. (2002). Generalized method of moments and macroeconomics. *Journal of Business & Economic Statistics* **20** 460–469. [MR1973798 https://doi.org/10.1198/073500102288618603](https://doi.org/10.1198/073500102288618603)
- JIA, Y. and LEE, L.-Y. (2007). Role of TRPV receptors in respiratory diseases. *Biochimica et Biophysica Acta (BBA)—Molecular Basis of Disease* **1772** 915–927.
- JOHNSON, D., BOYES, B., FIELDS, T., KOPKIN, R. and ORLANDO, R. (2013). Optimization of data-dependent acquisition parameters for coupling high-speed separations with LC-MS/MS for protein identifications. *Journal of Biomolecular Techniques* **24** 62–72. <https://doi.org/10.7171/jbt.13-2402-003>
- KAMINSKY, D. A. (2012). What does airway resistance tell us about lung function? *Respiratory Care* **57** 85–96; discussion 96–99. <https://doi.org/10.4187/respcare.01411>
- 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. *Statistical Science* **22** 523–539. [MR2420458 https://doi.org/10.1214/07-STS227](https://doi.org/10.1214/07-STS227)
- KARPIEVITCH, Y. V., POLPITIYA, A. D., ANDERSON, G. A., SMITH, R. D. and DABNEY, A. R. (2010). Liquid chromatography mass spectrometry-based proteomics: Biological and technical aspects. *The Annals of Applied Statistics* **4** 1797–1823.
- KIM, J.-Y. (2002). Limited information likelihood and Bayesian analysis. *Journal of Econometrics* **107** 175–193.
- KOEMAN, M., ENGEL, J., JANSEN, J. and BUYDENS, L. (2019). Critical comparison of methods for fault diagnosis in metabolomics data. *Scientific Reports* **9** 1123. <https://doi.org/10.1038/s41598-018-37494-7>
- LEE, S., SUN, W., WRIGHT, F. A. and ZOU, F. (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika* **104** 303–316. [MR3698255 https://doi.org/10.1093/biomet/asx018](https://doi.org/10.1093/biomet/asx018)
- LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* **3** 1724–1735.
- LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* **105** 18718–18723.
- LEEK, J. T., JOHNSON, W. E., PARKER, H. S., FERTIG, E. J., JAFFE, A. E., STOREY, J. D., ZHANG, Y. and TORRES, L. C. (2017). sva: Surrogate Variable Analysis. R package version 3.26.0.
- LI, C. and JIANG, W. (2016). On oracle property and asymptotic validity of Bayesian generalized method of moments. *Journal of Multivariate Analysis* **145** 132–147. [MR3459943 https://doi.org/10.1016/j.jmva.2015.12.009](https://doi.org/10.1016/j.jmva.2015.12.009)
- LIANG, K.-Y. and QIN, J. (2000). Regression analysis under non-standard situations: A pairwise pseudolikelihood approach. *Journal of the Royal Statistical Society: Series B* **62** 773–786. [MR1796291 https://doi.org/10.1111/1467-9868.00263](https://doi.org/10.1111/1467-9868.00263)

- LIU, X., SER, Z. and LOCASALE, J. W. (2014). Development and quantitative evaluation of a high-resolution metabolomics technology. *Analytical Chemistry* **86** 2175–2184.
- MCKENNAN, C. and NICOLAE, D. (2018). Estimating and accounting for unobserved covariates in high dimensional correlated data. Available at [arXiv:1808.05895v1](https://arxiv.org/abs/1808.05895v1).
- MCKENNAN, C. and NICOLAE, D. (2019). Accounting for unobserved covariates with varying degrees of estimability in high-dimensional biological data. *Biometrika* **106** 823–840. [MR4031201 https://doi.org/10.1093/biomet/asz037](https://doi.org/10.1093/biomet/asz037)
- MCKENNAN, C., OBER, C. and NICOLAE, D. (2020). Supplement to “Estimation and inference in metabolomics with nonrandom missing data and latent factors.” <https://doi.org/10.1214/20-AOAS1328SUPPA>, <https://doi.org/10.1214/20-AOAS1328SUPPB>
- O'BRIEN, J. J., GUNAWARDENA, H. P., PAULO, J. A., CHEN, X., IBRAHIM, J. G., GYGI, S. P. and QAQISH, B. F. (2018). The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *Ann. Appl. Stat.* **12** 2075–2095. [MR3875693 https://doi.org/10.1214/18-AOAS1144](https://doi.org/10.1214/18-AOAS1144)
- O'SULLIVAN, A., GIBNEY, M. J. and BRENNAN, L. (2010). Dietary intake patterns are reflected in metabolomic profiles: Potential role in dietary assessment studies. *Am. J. Clin. Nutr.* **93** 314–321.
- ONO, J. G., WORGALL, T. S. and WORGALL, S. (2015). Airway reactivity and sphingolipids-implications for childhood asthma. *Molecular and Cellular Pediatrics* **2** 13–13.
- OSTROUKHOVA, M., GOPLEN, N., KARIM, M. Z., MICHALEC, L., GUO, L., LIANG, Q. and ALAM, R. (2012). The role of low-level lactate production in airway inflammation in asthma. *Am. J. Physiol., Lung Cell. Mol. Physiol.* **302** L300–L307.
- PACOR, M. L., DI LORENZO, G., MARTINELLI, N., MANSUETO, P., RINI, G. B. and CORROCHER, R. (2004). Monosodium benzoate hypersensitivity in subjects with persistent rhinitis. *Allergy* **59** 192–197.
- PREMKUMAR, L. S. (2014). Transient receptor potential channels as targets for phytochemicals. *ACS Chemical Neuroscience* **5** 1117–1130.
- REINKE, S. N., GALLART-AYALA, H., GÓMEZ, C., CHECA, A., FAULAND, A., NAZ, S., KAMLEH, M. A., DJUKANOVIĆ, R., HINKS, T. S. C. et al. (2017). Metabolomics analysis identifies different metabotypes of asthma severity. *The European Respiratory Journal* **49** 1601740.
- SALERNO STEPHEN, J., MEHRMOHAMADI, M., LIBERTI, M. V., WAN, M., WELLS, M. T., BOOTH, J. G. and LOCASALE, J. W. (2017). RRMix: A method for simultaneous batch effect correction and analysis of metabolomics data in the absence of internal standards. *PLoS ONE* **12** e0179530.
- STOREY, J. D., BASS, A. J., DABNEY, A. and ROBINSON, D. (2015). qvalue: Q-value estimation for false discovery rate control. R package version 2.10.0.
- WANG, S., SHAO, J. and KIM, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statist. Sinica* **24** 1097–1116. [MR3241279](https://doi.org/10.1214/14-AOS1511)
- WANG, P., TANG, H., ZHANG, H., WHITEAKER, J., PAULOVICH, A. G. and MCINTOSH, M. (2006). Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 315–326.
- WANG, M., KONG, L., LI, Z. and ZHANG, L. (2016). Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples. *Stat. Med.* **35** 1706–1721. [MR3513479 https://doi.org/10.1002/sim.6817](https://doi.org/10.1002/sim.6817)
- WANG, J., ZHAO, Q., HASTIE, T. and OWEN, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *The Annals of Statistics* **45** 1863–1894. [MR3718155 https://doi.org/10.1214/16-AOS1511](https://doi.org/10.1214/16-AOS1511)
- WANG, J., WANG, P., HEDEKER, D. and CHEN, L. S. (2019). Using multivariate mixed-effects selection models for analyzing batch-processed proteomics data with non-ignorable missingness. *Biostatistics* **20** 648–665. [MR4019723 https://doi.org/10.1093/biostatistics/kxy022](https://doi.org/10.1093/biostatistics/kxy022)
- WEHRENS, R., HAGEMAN, J. A., VAN EEUWIJK, F., KOOKE, R., FLOOD, P. J., WIJNKER, E., KEURENTJES, J. J. B., LOMMEN, A., VAN EEKELEN, H. D. L. M. et al. (2016). Improved batch correction in untargeted MS-based metabolomics. *Metabolomics* **12** 88–88.
- XU, S., ROSS, C., RAEBEL, M. A., SHETTERLY, S., BLANCHETTE, C. and SMITH, D. (2010a). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value Health* **13** 273–277.
- XU, Y.-D., CUI, J.-M., WANG, Y., YIN, L.-M., GAO, C.-K., LIU, Y.-Y. and YANG, Y.-Q. (2010b). The early asthmatic response is associated with glycolysis, calcium binding and mitochondria activity as revealed by proteomic analysis in rats. *Respiratory Research* **11** 107.
- YIN, G. (2009). Bayesian generalized method of moments. *Bayesian Anal.* **4** 191–207. [MR2507358 https://doi.org/10.1214/09-BA407](https://doi.org/10.1214/09-BA407)

BAYESIAN VARIABLE SELECTION FOR SURVIVAL DATA USING INVERSE MOMENT PRIORS

BY AMIR NIKOOIENEJAD¹, WENYI WANG² AND VALEN E. JOHNSON³

¹Department of Global Statistical Sciences, Eli Lilly and Company, nikooienejad_amir@lilly.com

²Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center, wwang7@mdanderson.org

³Department of Statistics, Texas A&M University, vjohnson@stat.tamu.edu

Efficient variable selection in high-dimensional cancer genomic studies is critical for discovering genes associated with specific cancer types and for predicting response to treatment. Censored survival data is prevalent in such studies. In this article we introduce a Bayesian variable selection procedure that uses a mixture prior composed of a point mass at zero and an inverse moment prior in conjunction with the partial likelihood defined by the Cox proportional hazard model. The procedure is implemented in the R package BVSNLP, which supports parallel computing and uses a stochastic search method to explore the model space. Bayesian model averaging is used for prediction. The proposed algorithm provides better performance than other variable selection procedures in simulation studies and appears to provide more consistent variable selection when applied to actual genomic datasets.

REFERENCES

- ANTONIADIS, A., FRYZLEWICZ, P. and LETUÉ, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scand. J. Stat.* **37** 531–552. <https://doi.org/10.1111/j.1467-9469.2009.00685.x>
- BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. <https://doi.org/10.1214/009053604000000238>
- BENDER, R., AUGUSTIN, T. and BLETNER, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Stat. Med.* **24** 1713–1723. <https://doi.org/10.1002/sim.2059>
- BERGER, J. O., LISEO, B. and WOLPERT, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.* **14** 1–28. <https://doi.org/10.1214/ss/1009211803>
- BLANCHE, P., KATTAN, M. W. and GERDS, T. A. (2019). The c-index is not proper for the evaluation of t -year predicted risks. *Biostatistics* **20** 347–357. <https://doi.org/10.1093/biostatistics/kxy006>
- CANCER GENOME ATLAS RESEARCH NETWORK (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499** 43–49. <https://doi.org/10.1038/nature12222>
- CANCER GENOME ATLAS RESEARCH NETWORK (2016). Comprehensive molecular characterization of papillary renal-cell carcinoma. *N. Engl. J. Med.* **374** 135–145. <https://doi.org/10.1056/nejmoa1505917>
- COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220.
- COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data* **21**. CRC Press, Boca Raton, FL.
- ELSEVIER (2018). PathwayStudio[®] pathwaystudio.com. Elsevier.
- FAN, J., FENG, Y. and WU, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications—a Festschrift for Lawrence D. Brown*. *Inst. Math. Stat. (IMS) Collect.* **6** 70–86. IMS, Beachwood, OH. <https://doi.org/10.1214/10-imscol606>
- FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. <https://doi.org/10.1214/aos/1015362185>
- FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. <https://doi.org/10.1111/j.1467-9868.2008.00674.x>
- FARAGGI, D. and SIMON, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics* **54** 1475–1485. <https://doi.org/10.2307/2533672>
- FENNER, A. (2016). Kidney cancer: AR promotes RCC via lncRNA interaction. *Nat. Rev. Urol.* **13** 242. <https://doi.org/10.1038/nrurol.2016.61>

Key words and phrases. Bayesian variable selection, nonlocal prior, high-dimensional data, survival data analysis, Cox proportional hazard model, cancer genomics.

- 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>
- GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.
- GERDS, T. A. and SCHUMACHER, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom. J.* **48** 1029–1040. <https://doi.org/10.1002/bimj.200610301>
- GHOSH, J. K. (1988). *Statistical Information and Likelihood: A Collection of Critical Essays by Dr. D. Basu. Lect. Notes in Statist* **45**.
- HARRELL, F. E. JR., CALIFF, R. M., PRYOR, D. B., LEE, K. L., ROSATI, R. A. et al. (1982). Evaluating the yield of medical tests. *JAMA* **247** 2543–2546. <https://doi.org/10.1001/jama.247.18.2543>
- HATA, S., ISE, K., AZMAHANI, A., KONOSU-FUKAYA, S., MCNAMARA, K. M., FUJISHIMA, F., SHIMADA, K., MITSUZUKA, K., ARAI, Y. et al. (2017). Expression of AR, 5 α R1 and 5 α R2 in bladder urothelial carcinoma and relationship to clinicopathological factors. *Life Sci.* **190** 15–20. <https://doi.org/10.1016/j.lfs.2017.09.029>
- HEAGERTY, P. J., LUMLEY, T. and PEPE, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56** 337–344. <https://doi.org/10.1111/j.0006-341x.2000.00337.x>
- HELD, L., GRAVESTOCK, I. and SABANÉS BOVÉ, D. (2016). Objective Bayesian model selection for Cox regression. *Stat. Med.* **35** 5376–5390. <https://doi.org/10.1002/sim.7089>
- IBRAHIM, J. G., CHEN, M.-H. and MACÉACHERN, S. N. (1999). Bayesian variable selection for proportional hazards models. *Canad. J. Statist.* **27** 701–717. <https://doi.org/10.2307/3316126>
- JOHNSON, V. E. (2005). Bayes factors based on test statistics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 689–701. <https://doi.org/10.1111/j.1467-9868.2005.00521.x>
- JOHNSON, V. E. and ROSSELL, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 143–170. <https://doi.org/10.1111/j.1467-9868.2009.00730.x>
- JOHNSON, V. E. and ROSSELL, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. <https://doi.org/10.1080/01621459.2012.682536>
- 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.
- LIU, D. C. and NOCEDAL, J. (1989). On the limited memory BFGS method for large scale optimization. *Math. Program.* **45** 503–528. <https://doi.org/10.1007/BF01589116>
- MALUMBRES, M. and BARBACID, M. (2009). Cell cycle, CDKs and cancer: A changing paradigm. *Nat. Rev. Cancer* **9** 153–166. <https://doi.org/10.1038/nrc2602>
- MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. <https://doi.org/10.1111/j.1467-9868.2010.00740.x>
- MOGENSEN, U. B., ISHWARAN, H. and GERDS, T. A. (2012). Evaluating random forests for survival analysis using prediction error curves. *J. Stat. Softw.* **50** 1–23. <https://doi.org/10.18637/jss.v050.i11>
- MORRIS, T. P., WHITE, I. R. and CROWTHER, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* **38** 2074–2102. <https://doi.org/10.1002/sim.8086>
- NIKITIN, A., EGOROV, S., DARASELIA, N. and MAZO, I. (2003). Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics* **19** 2155–2157. <https://doi.org/10.1093/bioinformatics/btg290>
- NIKOOIENEJAD, A., WANG, W. and JOHNSON, V. E. (2016). Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. *Bioinformatics* **32** 1338–1345. <https://doi.org/10.1093/bioinformatics/btv764>
- NIKOOIENEJAD, A., WANG, W. and JOHNSON, V. E. (2020). Supplement to “Bayesian variable selection for survival data using inverse moment priors.” <https://doi.org/10.1214/20-AOAS1325SUPP>
- SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. <https://doi.org/10.1214/10-AOS792>
- SHA, N., TADESSE, M. G. and VANNUCCI, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics* **22** 2262–2268. <https://doi.org/10.1093/bioinformatics/btl362>
- SHIN, M., BHATTACHARYA, A. and JOHNSON, V. E. (2018). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statist. Sinica* **28** 1053–1078. <https://doi.org/10.5705/ss.202016.0167>
- SILL, M., HIELSCHER, T., BECKER, N. and ZUCKNICK, M. (2014). c060: Extended inference with Lasso and elastic-net regularized Cox and generalized linear models. *J. Stat. Softw.* **62** 1–22. <https://doi.org/10.18637/jss.v062.i05>
- SUZUKI, H., OUCHIDA, M., YAMAMOTO, H., YANO, M., TOYOOKA, S., AOE, M., SHIMIZU, N., DATE, H. and SHIMIZU, K. (2008). Decreased expression of the SIN3A gene, a candidate tumor suppressor located at the prevalent allelic loss region 15q23 in non-small cell lung cancer. *Lung Cancer* **59** 24–31. <https://doi.org/10.1016/j.lungcan.2007.08.002>

- TIBSHIRANI, R. et al. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395. [https://doi.org/10.1002/\(sici\)1097-0258\(19970228\)16:4<385::aid-sim380>3.0.co;2-3](https://doi.org/10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3)
- UNO, H., CAI, T., TIAN, L. and WEI, L. J. (2007). Evaluating prediction rules for t -year survivors with censored regression models. *J. Amer. Statist. Assoc.* **102** 527–537. <https://doi.org/10.1198/016214507000000149>
- VAN HOUWELINGEN, H. C. and PUTTER, H. (2011). *Dynamic Prediction in Clinical Survival Analysis. Monographs on Statistics and Applied Probability* **123**. CRC Press, Boca Raton, FL. <https://doi.org/10.1201/b11311>
- WANG, Z., MORRIS, J. S., CAO, S., AHN, J., LIU, R., TYEKUCHEVA, S., LI, B., LU, W., TANG, X. et al. (2017). Transcriptome deconvolution of heterogeneous tumor samples with immune infiltration. *BioRxiv*. <https://doi.org/10.1101/146795>
- ZHANG, H. H. and LU, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94** 691–703. <https://doi.org/10.1093/biomet/asm037>

EVIDENCE FACTORS IN A CASE-CONTROL STUDY WITH APPLICATION TO THE EFFECT OF FLEXIBLE SIGMOIDOSCOPY SCREENING ON COLORECTAL CANCER

BY BIKRAM KARMAKAR¹, CHYKE A. DOUBENI² AND DYLAN S. SMALL³

¹*Department of Statistics, College of Liberal Arts and Sciences, University of Florida, bkarmakar@ufl.edu*

²*Center for Health Equity and Community Engagement Research, Mayo Clinic, Doubeni.Chyke@mayo.edu*

³*Department of Statistics, The Wharton School, University of Pennsylvania, dsmall@wharton.upenn.edu*

As in any observational study, in a case-control study a primary concern is potential unmeasured confounders. Bias, due to unmeasured confounders, can result in a false discovery of an apparent treatment effect when there is none. Replication of an observational study, which tries to provide multiple analyses of the data where the biases affecting each analysis are thought to be different, is one way to strengthen the evidence from an observational study. Evidence factors allow for internal replication by testing a hypothesis using multiple comparisons in a way that the comparisons yield independent evidence and differ in the sources of potential bias. We construct evidence factors in a case-control study in which there are two types of cases, “narrow” cases which are thought to be potentially more affected by the exposure and “marginal” cases which are thought to have more heterogeneous causes. We develop and study an inference procedure for using such evidence factors and apply it to a study of the effect of sigmoidoscopy screening on colorectal cancer.

REFERENCES

- ACHESON, E. D. (1979). Comment on “The evolving case-control study.” *J. Chronic Dis.* **32** 28–29.
- ATKIN, W. S., EDWARDS, R., KARLJ-HANS, I. et al. (2010). Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: A multicentre randomized controlled trial. *Lancet* **375** 1624–1633.
- BAZZANO, L. A., HE, J., MUNTNER, P., VUPPUTURI, S. and WHELTON, P. K. (2003). Relationship between cigarette smoking and novel risk factors for cardiovascular disease in the United States. *Ann. Intern. Med.* **138** 891–897.
- BECKER, B. J. (1994). Combining significance levels. In *The Handbook of Research Synthesis* 215–230. Russell Sage Foundation, Thousand Oaks, CA.
- BIZZOZERO, O. J., JOHNSON, K. G. and CIOCCO, A. (1966). Radiation related leukemia in Hiroshima and Nagasaki, 1946–1964. *N. Engl. J. Med.* **274** 1095–1101.
- COCHRAN, W. G. (1965). The planning of observational studies of human populations. *J. Roy. Statist. Soc. Ser. A* **128** 134–155. Reprinted in *Readings in Economic Statistics and Econometrics* (A. Zellner, ed.), Little Brown, Boston, MA, pp. 11–36 (1968).
- COHEN, J. C., KISS, R. S., PERTSEMLIDIS, A., KOTOWSKI, I. K., GRAHAM, R., KIM GARCIA, C. and HOBBS, H. H. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent non-sense mutation in PCSK9. *Nat. Genet.* **37** 161–165.
- COLE, P. (1979). The evolving case-control study. *J. Chronic Dis.* **32** 15–27. [https://doi.org/10.1016/0021-9681\(79\)90006-7](https://doi.org/10.1016/0021-9681(79)90006-7)
- DOUBENI, C. A., MAJOR, J. M., LAIYEMO, A. O., SCHOOTMAN, M., ZAUBER, A. G., HOLLENBECK, A. R., SINHA, R. and ALLISON, J. (2012). Contribution of behavioral risk factors and obesity to socioeconomic differences in colorectal cancer incidence. *J. Natl. Cancer Inst.* **104** 1353–1362.
- DOUBENI, C. A., CORLEY, D. A., QUINN, V. P., JENSEN, C. D., ZAUBER, A. G., GOODMAN, M., JOHNSON, J. R., MEHTA, S. J., BECERRA, T. A. et al. (2018). Effectiveness of screening colonoscopy in reducing the risk of death from right and left colon cancer: A large community-based study. *Gut* **67** 291–298. <https://doi.org/10.1136/gutjnl-2016-312712>

- ELDRIDGE, R. C., DOUBENI, C. A., FLETCHER, R. H., ZAUBER, A. G., CORLEY, D. A., DORIA-ROSE, V. P. and GOODMAN, M. (2013). Uncontrolled confounding in studies of screening effectiveness: An example of colonoscopy. *J. Med. Screen.* **20** 198–207. <https://doi.org/10.1177/0969141313508282>
- FISHER, R. A. (1932). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.
- GAIST, D., JEPPESEN, U., ANDERSEN, M., GARCÍA RODRÍGUEZ, A. L., HALLAS, J. and SINDRUP, H. S. (2002). Statins and risk of polyneuropathy: A case-control study. *Neurology* **58** 1333–1337.
- GASTWIRTH, J. L., KRIEGER, A. M. and ROSENBAUM, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika* **85** 907–920.
- GOODMAN, M., FLETCHER, R. H., DORIA-ROSE, V. P., JENSEN, C. D., ZEBROWSKI, A. M., BECERRA, T. A., QUINN, V. P., ZAUBER, A. G., CORLEY, D. A. et al. (2015). Observational methods to assess the effectiveness of screening colonoscopy in reducing right colon cancer mortality risk: SCOLAR. *J. Comp. Eff. Res.* **4** 541–551.
- HANSEN, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist. Assoc.* **99** 609–618. [MR2086387 https://doi.org/10.1198/016214504000000647](https://doi.org/10.1198/016214504000000647)
- HANSEN, B. B. and KLOPFER, S. O. (2006). Optimal full matching and related designs via network flows. *J. Comput. Graph. Statist.* **15** 609–627. [MR2280151 https://doi.org/10.1198/106186006X137047](https://doi.org/10.1198/106186006X137047)
- HILL, A. B. (1965). The environment and disease: Association or causation? *Proc. R. Soc. Med.* **58** 295–300.
- HSU, J. Y., SMALL, D. S. and ROSENBAUM, P. R. (2013). Effect modification and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **108** 135–148. [MR3174608 https://doi.org/10.1080/01621459.2012.742018](https://doi.org/10.1080/01621459.2012.742018)
- JOSEPH, D. A., MEESTER, R. G. S., ZAUBER, A. G., MANNINEN, D. L., WINGES, L., DONG, F. B., PEAKER, B. and VAN BALLEGOIJEN, M. (2016). Colorectal cancer screening: Estimated future colonoscopy need and current volume and capacity. *Cancer* **122** 2479–2486.
- KARMAKAR, B., FRENCH, B. and SMALL, D. S. (2019). Integrating the evidence from evidence factors in observational studies. *Biometrika* **106** 353–367. [MR3949308 https://doi.org/10.1093/biomet/asz003](https://doi.org/10.1093/biomet/asz003)
- KARMAKAR, B., SMALL, D. S. and ROSENBAUM, P. R. (2020). Using evidence factors to clarify exposure biomarkers. *Am. J. Epidemiol.* To appear. <https://doi.org/10.1093/aje/kwz263>
- KEOGH, R. H. and COX, D. R. (2014). *Case-Control Studies*. Institute of Mathematical Statistics (IMS) Monographs **4**. Cambridge Univ. Press, Cambridge. [MR3443808 https://doi.org/10.1017/CBO9781139094757](https://doi.org/10.1017/CBO9781139094757)
- LEWIS, E. B. (1963). Leukemia, multiple myeloma, and aplastic anemia in American radiologists. *Science* **142** 1492–1494.
- LIPTAK, T. (1958). On the combination of independent tests. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* **3** 171–197.
- MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. [MR0468056 https://doi.org/10.1093/biomet/63.3.655](https://doi.org/10.1093/biomet/63.3.655)
- MING, K. and ROSENBAUM, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* **56** 118–124.
- MISSIAGLIA, E., JACOBS, B., D'ARIO, G., NARZO, A. F. D., SONESON, C., BUDINSKA, E., POPOVICI, V., VECCHIONE, L., GERSTER, S. et al. (2014). Distal and proximal colon cancers differ in terms of molecular, pathological, and clinical features. *Ann. Oncol.* **25** 1995–2001. <https://doi.org/10.1093/annonc/mdu275>
- NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Ann. Agric. Sci.* **10** 1–51 (in Polish). [Reprinted in English with discussion by T. Speed and D. B. Rubin in *Statist. Sci.* **5** (1990) 463–480.] [MR1092986](https://doi.org/10.1093/biomet/asz003)
- PIMENTEL, S. D., YOON, F. and KEELE, L. (2015). Variable-ratio matching with fine balance in a study of the Peer Health Exchange. *Stat. Med.* **34** 4070–4082. [MR3431322 https://doi.org/10.1002/sim.6593](https://doi.org/10.1002/sim.6593)
- POPPER, K. R. (1959). *The Logic of Scientific Discovery*. Hutchinson and Co., Ltd., London. [MR0107593](https://doi.org/10.1093/biomet/asz003)
- U. S. PREVENTIVE SERVICES TASK FORCE, BIBBINS-DOMINGO, K., GROSSMAN, D. C. et al. (2016). Screening for colorectal cancer: US preventive services task force recommendation statement. *J. Am. Med. Assoc.* **315** 2564–2575.
- ROSENBAUM, P. R. (1991). Sensitivity analysis for matched case-control studies. *Biometrics* **47** 87–100. [MR1108691 https://doi.org/10.2307/2532498](https://doi.org/10.2307/2532498)
- ROSENBAUM, P. R. (2001). Replicating effects and biases. *Amer. Statist.* **55** 223–227. [MR1963397 https://doi.org/10.1198/000313001317098220](https://doi.org/10.1198/000313001317098220)
- ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR1899138 https://doi.org/10.1007/978-1-4757-3692-2](https://doi.org/10.1007/978-1-4757-3692-2)
- ROSENBAUM, P. R. (2010). Evidence factors in observational studies. *Biometrika* **97** 333–345. [MR2650742 https://doi.org/10.1093/biomet/asq019](https://doi.org/10.1093/biomet/asq019)
- ROSENBAUM, P. R. (2011). Some approximate evidence factors in observational studies. *J. Amer. Statist. Assoc.* **106** 285–295. [MR2816721 https://doi.org/10.1198/jasa.2011.tm10422](https://doi.org/10.1198/jasa.2011.tm10422)
- ROSENBAUM, P. R. (2017). The general structure of evidence factors in observational studies. *Statist. Sci.* **32** 514–530. [MR3730520 https://doi.org/10.1214/17-STS621](https://doi.org/10.1214/17-STS621)

- ROSENBAUM, P. R., ROSS, R. N. and SILBER, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Amer. Statist. Assoc.* **102** 75–83. [MR2345534 https://doi.org/10.1198/016214506000001059](https://doi.org/10.1198/016214506000001059)
- ROSENBAUM, P. R. and SILBER, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *J. Amer. Statist. Assoc.* **104** 1398–1405. [MR2750570 https://doi.org/10.1198/jasa.2009.tm08470](https://doi.org/10.1198/jasa.2009.tm08470)
- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.
- SEGNAN, N., ARMAROLI, P., BONELLI, L. et al. (2011). Once-only sigmoidoscopy in colorectal cancer screening: Follow-up findings of the Italian Randomized Control Trial—SCORE. *J. Natl. Cancer Inst.* **103** 1310–1322.
- SELBY, J. V., FRIEDMAN, G. D., QUESENBERRY, C. P. and WEISS, N. (1992). A case-control study of screening sigmoidoscopy and mortality from colorectal cancer. *N. Engl. J. Med.* **326** 653–657.
- 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](https://doi.org/10.1007/978-0-387-34675-5)
- SMALL, D. S., CHENG, J., HALLORAN, M. E. and ROSENBAUM, P. R. (2013). Case definition and design sensitivity. *J. Amer. Statist. Assoc.* **108** 1457–1468. [MR3174721 https://doi.org/10.1080/01621459.2013.820660](https://doi.org/10.1080/01621459.2013.820660)
- SMALL, D. S., TAYLOR, T. E., POSTELS, D. G., BEARE, N. A., CHENG, J., MACCORMICK, I. J. and SEYDEL, K. B. (2017). Evidence from a natural experiment that malaria parasitemia is pathogenic in retinopathy-negative cerebral malaria. *eLife* **6**. <https://doi.org/10.7554/eLife.23699>
- TAYLOR, T. E., FU, W. J., CARR, R. A., WHITTEN, R. O., MUELLER, J. S., FOSIKO, N. G., LEWALLEN, S., LIOMBA, N. G., MOLYNEUX, M. E. et al. (2004). Differentiating the pathologies of cerebral malaria by postmortem parasite counts. *Nat. Med.* **10** 143–145.
- WITTGENSTEIN, L. (1958). *Philosophical Investigations*, 2nd ed. The Macmillan Co., New York. [MR0078292 https://doi.org/10.1080/01621459.2013.820660](https://doi.org/10.1080/01621459.2013.820660)
- ZAYKIN, D., ZHIVOTOVSKY, L. A., WESTFALL, P. and WEIR, B. (2002). Truncated product method for combining p -values. *Genet. Epidemiol.* **22** 170–185.
- ZHANG, K., SMALL, D. S., LORCH, S., SRINIVAS, S. and ROSENBAUM, P. R. (2011). Using split samples and evidence factors in an observational study of neonatal outcomes. *J. Amer. Statist. Assoc.* **106** 511–524. [MR2847966 https://doi.org/10.1198/jasa.2011.ap10604](https://doi.org/10.1198/jasa.2011.ap10604)
- ZUBIZARRETA, J. R., NEUMAN, M., SILBER, J. H. and ROSENBAUM, P. R. (2012). Contrasting evidence within and between institutions that provide treatment in an observational study of alternative forms of anesthesia. *J. Amer. Statist. Assoc.* **107** 901–915. [MR3010879 https://doi.org/10.1080/01621459.2012.682533](https://doi.org/10.1080/01621459.2012.682533)
- R CORE TEAM (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. Available at <http://www.R-project.org>.

A CAUSAL EXPOSURE RESPONSE FUNCTION WITH LOCAL ADJUSTMENT FOR CONFOUNDING: ESTIMATING HEALTH EFFECTS OF EXPOSURE TO LOW LEVELS OF AMBIENT FINE PARTICULATE MATTER

BY GEORGIA PAPADOGEORGOU¹ AND FRANCESCA DOMINICI²

¹*Department of Statistical Science, Duke University, gp118@duke.edu*

²*Department of Biostatistics, Harvard T.H. Chan School of Public Health, fdominic@hsph.harvard.edu*

In the last two decades ambient levels of air pollution have declined substantially. At the same time the Clean Air Act mandates that the National Ambient Air Quality Standards (NAAQS) must be routinely assessed to protect populations based on the latest science. Therefore, researchers should continue to address the following question: is exposure to levels of air pollution below the NAAQS harmful to human health? Furthermore, the contentious nature surrounding environmental regulations urges us to cast this question within a causal inference framework. Several parametric and semiparametric regression approaches have been used to estimate the exposure-response (ER) curve between long-term exposure to ambient air pollution concentrations and health outcomes. However, most of the existing approaches are not formulated within a formal framework for causal inference, adjust for the same set of potential confounders across all levels of exposure and do not account for model uncertainty regarding covariate selection and the shape of the ER.

In this paper we introduce a Bayesian framework for the estimation of a causal ER curve called LERCA (Local Exposure Response Confounding Adjustment), which: (a) allows for different confounders *and* different strength of confounding at the different exposure levels, and (b) propagates model uncertainty regarding confounders' selection and the shape of the ER. Importantly, LERCA provides a principled way of assessing the observed covariates' confounding importance at different exposure levels, providing researchers with important information regarding the set of variables to measure and adjust for in regression models. Using simulation studies, we show that state-of-the-art approaches perform poorly in estimating the ER curve in the presence of local confounding.

LERCA is used to evaluate the relationship between long-term exposure to ambient PM_{2.5}, a key regulated pollutant, and cardiovascular hospitalizations for 5,362 zip codes in the continental U.S. and located near a pollution monitoring site, while adjusting for a potentially varying set of confounders across the exposure range. Our data set includes rich health, weather, demographic and pollution information for the years of 2011–2013. The estimated exposure-response curve is increasingly indicating that higher ambient concentrations lead to higher cardiovascular hospitalization rates, and ambient PM_{2.5} was estimated to lead to an increase in cardiovascular hospitalization rates when focusing at the low-exposure range. Our results indicate that there is no threshold for the effect of PM_{2.5} on cardiovascular hospitalizations.

REFERENCES

- ANDROUTSOPOULOS, V. P., HERNANDEZ, A. F., LIESIVUORI, J. and TSATSAKIS, A. M. (2013). A mechanistic overview of health associated effects of low levels of organochlorine and organophosphorous pesticides. *Toxicology* **307** 89–94. <https://doi.org/10.1016/j.tox.2012.09.011>

Key words and phrases. Air pollution, cardiovascular hospitalizations, causal inference, exposure response function, local confounding, low-exposure levels, particulate matter.

- ANTONELLI, J., PARMIGIANI, G. and DOMINICI, F. (2019). High-dimensional confounding adjustment using continuous spike and slab priors. *Bayesian Anal.* **14** 825–848. [MR3960772 https://doi.org/10.1214/18-BA1131](https://doi.org/10.1214/18-BA1131)
- ANTONELLI, J., ZIGLER, C. and DOMINICI, F. (2017). Guided Bayesian imputation to adjust for confounding when combining heterogeneous data sources in comparative effectiveness research. *Biostatistics* **18** 553–568. [MR3799594 https://doi.org/10.1093/biostatistics/kxx003](https://doi.org/10.1093/biostatistics/kxx003)
- BABB, J., ROGATKO, A. and ZACKS, S. (1998). Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Stat. Med.* **17** 1103–1120. [https://doi.org/10.1002/\(sici\)1097-0258\(19980530\)17:10<1103::aid-sim793>3.0.co;2-9](https://doi.org/10.1002/(sici)1097-0258(19980530)17:10<1103::aid-sim793>3.0.co;2-9)
- BELL, M. L., PENG, R. D. and DOMINICI, F. (2006). The exposure-response curve for ozone and risk of mortality and the adequacy of current ozone regulations. *Environ. Health Perspect.* **114** 532–536.
- BERGER, R. E., RAMASWAMI, R., SOLOMON, C. G. and DRAZEN, J. M. (2017). Air pollution still kills. *N. Engl. J. Med.* **376** 2591–2592.
- BIA, M., FLORES, C. A., FLORES-LAGUNES, A. and MATTEI, A. (2014). A Stata package for the application of semiparametric estimators of dose-response functions. *Stata J.* **14** 580–604.
- BURNETT, R. T., ARDEN POPE, C., EZZATI, M., OLIVES, C., LIM, S. S., MEHTA, S., SHIN, H. H., SINGH, G., HUBBELL, B. et al. (2014). An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. *Environ. Health Perspect.* **122** 397–403. <https://doi.org/10.1289/ehp.1307049>
- CEFALU, M., DOMINICI, F., ARVOLD, N. and PARMIGIANI, G. (2017). Model averaged double robust estimation. *Biometrics* **73** 410–421. [MR3665958 https://doi.org/10.1111/biom.12622](https://doi.org/10.1111/biom.12622)
- CHIDO, L. M., JACOBSON, S. W. and JACOBSON, J. L. (2004). Neurodevelopmental effects of postnatal lead exposure at very low levels. *Neurotoxicol. Teratol.* **26** 359–371.
- CHUA, K. L., SOH, S. E., MA, S. and LEE, B. W. (2009). Pediatric asthma mortality and hospitalization trends across Asia Pacific. *World Allergy Organ. J.* **2** 77–82.
- CLAYTON, C. A., PERRITT, R. L., PELLIZZARI, E. D., THOMAS, K. W., WHITMORE, R. W., WALLACE, L. A., OZKAYNAK, H. and SPENGLER, J. D. (1993). Particle Total Exposure Assessment Methodology (PTEAM) study: Distributions of aerosol and elemental concentrations in personal, indoor, and outdoor air samples in a southern California community. *J. Expo. Anal. Environ. Epidemiol.* **3** 227–250.
- CROUSE, D. L., PETERS, P. A., HYSTAD, P., BROOK, J. R., VAN DONKELAAR, A., MARTIN, R. V., VILLENEUVE, P. J., JERRETT, M., GOLDBERG, M. S. et al. (2015). Ambient PM_{2.5}, O₃, and NO₂ exposures and associations with mortality over 16 years of follow-up in the Canadian Census Health and Environment Cohort (CanCHEC). *Environ. Health Perspect.* **123** 1180–1186. <https://doi.org/10.1289/ehp.1409276>
- CROUSE, D. L., PHILIP, S., DONKELAAR, A. V., MARTIN, R. V., JESSIMAN, B., PETERS, P. A., WEICHTHAL, S., BROOK, J. R., HUBBELL, B. et al. (2016). A new method to jointly estimate the mortality risk of long-term exposure to fine particulate matter and its components. *Sci. Rep.* **6** Art. ID 18916. <https://doi.org/10.1038/srep18916>
- DANIELS, M. J., DOMINICI, F., SAMET, J. M. and ZEGER, S. L. (2000). Estimating particulate matter-mortality dose-response curves and threshold levels: An analysis of daily time-series for the 20 largest US cities. *Am. J. Epidemiol.* **152** 397–406. <https://doi.org/10.1093/aje/152.5.397>
- DANIELS, M. J., DOMINICI, F., ZEGER, S. L. and SAMET, J. M. (2004). The national morbidity, mortality, and air pollution study. Part III: PM₁₀ concentration-response curves and thresholds for the 20 largest US cities. Research Report 94, Health Effects Institute.
- DE LUNA, X., WAERNBAUM, I. and RICHARDSON, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika* **98** 861–875. [MR2860329 https://doi.org/10.1093/biomet/asr041](https://doi.org/10.1093/biomet/asr041)
- DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998). Automatic Bayesian curve fitting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 333–350. [MR1616029 https://doi.org/10.1111/1467-9868.00128](https://doi.org/10.1111/1467-9868.00128)
- DEVOS, S., COX, B., LIER, T. V., NAWROT, T. S. and PUTMAN, K. (2016). Effect of the shape of the exposure-response function on estimated hospital costs in a study on non-elective pneumonia hospitalizations related to particulate matter. *Environ. Int.* **94** 525–530.
- DEVRIES, R., KRIEBEL, D. and SAMA, S. (2016). Low level air pollution and exacerbation of existing COPD: A case crossover analysis. *Environ. Health* **15** Art. ID 98. <https://doi.org/10.1186/s12940-016-0179-z>
- DI, Q., DAI, L., WANG, Y., ZANOBBETTI, A., CHOIRAT, C., SCHWARTZ, J. D. and DOMINICI, F. (2017a). Association of short-term exposure to air pollution with mortality in older adults. *J. Am. Med. Assoc.* **318** 2446–2456. <https://doi.org/10.1001/jama.2017.17923>
- DI, Q., WANG, Y., ZABONETTI, A., WANG, Y., KOUTRAKIS, P., CHOIRAT, C., DOMINICI, F. and SCHWARTZ, J. D. (2017b). Air pollution and mortality in the medicare population. *N. Engl. J. Med.* **376** 2513–2522. <https://doi.org/10.1056/NEJMoa1702747>
- DIMATTEO, I., GENOVESE, C. R. and KASS, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88** 1055–1071. [MR1872219 https://doi.org/10.1093/biomet/88.4.1055](https://doi.org/10.1093/biomet/88.4.1055)

- DOCKERY, D. W. and SPENGLER, J. D. (1981). Personal exposure to respirable particulates and sulfates. *J. Air Pollut. Control Assoc.* **31** 153–159. <https://doi.org/10.1080/00022470.1981.10465205>
- DOMINICI, F., DANIELS, M., ZEGER, S. L. and SAMET, J. M. (2002). Air pollution and mortality: Estimating regional and national dose-response relationships. *J. Amer. Statist. Assoc.* **97** 100–111. [MR1963390 https://doi.org/10.1198/016214502753479266](https://doi.org/10.1198/016214502753479266)
- EFTIM, S. E., SAMET, J. M., JANES, H., McDERMOTT, A. and DOMINICI, F. (2008). Fine particulate matter and mortality: A comparison of the six cities and American Cancer Society cohorts with a medicare cohort. *Epidemiology* **19** 209–216.
- FAZEL, R., KRUMHOLZ, H. M., WANG, Y., ROSS, J. S., CHEN, J., TING, H. H., SHAH, N. D., NASIR, K., EINSTEIN, A. J. et al. (2009). Exposure to low-dose ionizing radiation from medical imaging procedures. *N. Engl. J. Med.* **361** 849–857.
- GARCIA, C. A., YAP, P., PARK, H. and WELLER, B. L. (2016). Association of long-term PM_{2.5} exposure with mortality using different air pollution exposure models: Impacts in rural and urban California. *Int. J. Environ. Health Res.* **26** 145–157. <https://doi.org/10.1080/09603123.2015.1061113>
- 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](https://doi.org/10.1007/s11222-013-9416-2)
- GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–511.
- FLORES, C. A., NEUMANN, A., GONZALEZ, T. C. and GONZALEZ, A. (2012). Estimating the effects of length of exposure to instruction in a training program: The case of job corps. *Rev. Econ. Stat.* **94** 153–171.
- GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. [MR1380810 https://doi.org/10.1093/biomet/82.4.711](https://doi.org/10.1093/biomet/82.4.711)
- HART, J. E., LIAO, X., HONG, B., PUETT, R. C., YANOSKY, J. D., SUH, H., KIOUMOURTZOGLOU, M., SPIEGELMAN, D. and LADEN, F. (2015). The association of long-term exposure to PM_{2.5} on all-cause mortality in the Nurses' Health Study and the impact of measurement-error correction. *Environ. Health* **14** Art. ID 38. <https://doi.org/10.1186/s12940-015-0027-6>
- HASTIE, T. (2017). gam: Generalized additive models.
- HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models. *Statist. Sci.* **1** 297–318. [MR0858512](https://doi.org/10.1198/0162145000001187)
- HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. [MR1229881](https://doi.org/10.1198/0162145000001187)
- HIRANO, K. and IMBENS, G. W. (2004). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. Wiley Ser. Probab. Stat. 73–84. Wiley, Chichester. [MR2134803 https://doi.org/10.1002/0470090456.ch7](https://doi.org/10.1002/0470090456.ch7)
- IMAI, K. and VAN DYK, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* **99** 854–866. [MR2090918 https://doi.org/10.1198/016214504000001187](https://doi.org/10.1198/016214504000001187)
- JERRETT, M., TURNER, M. C., BECKERMAN, B. S., III, C. A. P. and DONKELAAR, A. V. (2017). Comparing the health effects of ambient particulate matter estimated using ground-based versus remote sensing exposure estimates. *Environ. Health Perspect.* **125** 552–559.
- JOSHUA, S. C. and GARBER, N. J. (1990). Estimating truck accident rate and involvements using linear and Poisson regression models. *Transp. Plann. Technol.* **15** 41–58.
- JUSKO, T. A., HENDERSON, C. R., LANPHEAR, B. P., CORY-SLECHTA, D. A., PARSONS, P. J. and CANFIELD, R. L. (2008). Blood lead concentrations < 10 µg/dL and child intelligence at 6 years of age. *Environ. Health Perspect.* **116** 243–248.
- KENNEDY, E. H., MA, Z., MCHUGH, M. D. and SMALL, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1229–1245. [MR3689316 https://doi.org/10.1111/rssb.12212](https://doi.org/10.1111/rssb.12212)
- LEE, M., KOUTRAKIS, P., COULL, B. A., KLOOG, I. and SCHWARTZ, J. D. (2016). Acute effect of fine particulate matter on mortality in three southeastern states 2007–2011. *J. Expo. Sci. Environ. Epidemiol.* **26** 173–179. <https://doi.org/10.1038/jes.2015.47>
- LIM, C. C., HAYES, R. B., AHN, J., SHAO, Y., SILVERMAN, D. T., JONES, R. R., GARCIA, C. and THURSTON, G. D. (2018). Association between long-term exposure to ambient air pollution and diabetes mortality in the US. *Environ. Res.* **165** 330–336. <https://doi.org/10.1016/j.envres.2018.04.011>
- LIU-SMITH, F., FARHAT, A. M., ARCE, A., ZIOGAS, A., TAYLOR, T., WANG, Z., YOURK, V., LIU, J., WU, J. et al. (2016). Gender differences in the association between the incidence of cutaneous melanoma and geographic UV exposure. *J. Am. Acad. Dermatol.* **76** 499–505. <https://doi.org/10.1016/j.jaad.2016.08.027>
- MAKAR, M., ANTONELLI, J., DI, Q., CUTLER, D. and SCHWARTZ, J. (2018). Estimating the causal effect of fine particulate matter levels on death and hospitalization: Are levels below the safety standards harmful?. *Epidemiology* **28** 627–634. <https://doi.org/10.1097/EDE.0000000000000690>

- MENG, Q. Y., SPECTOR, D., COLOME, S. and TURPIN, B. (2009). Determinants of indoor and personal exposure to PM(2.5) of indoor and outdoor origin during the RIOPA study. *Atmos. Environ.* **43** 5750–5758. <https://doi.org/10.1016/j.atmosenv.2009.07.066>
- MOHAMEDSHAH, Y. M., PANIATI, J. F. and HOBEIKA, A. G. (1993). Truck accident models for interstates and two-lane rural roads. Transportation Research Record 1407, TRB, National Research Council.
- NASARI, M. M., SZYSZKOWICZ, M., CHEN, H., CROUSE, D., TURNER, M. C., JERRETT, M., POPE, C. A., HUBBELL, B., FANN, N. et al. (2016). A class of non-linear exposure-response models suitable for health impact assessment applicable to large cohort studies of ambient air pollution. *Air Qual. Atmos. Health* **9** 961–972. <https://doi.org/10.1007/s11869-016-0398-z>
- NATIONAL RESEARCH COUNCIL (2006). *Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2*. The National Academies Press, Washington, DC.
- PAPADOGEORGOU, G. and DOMINICI, F. (2020). Supplement to “A causal exposure response function with local adjustment for confounding: Estimating health effects of exposure to low levels of ambient fine particulate matter.” <https://doi.org/10.1214/20-AOAS1330SUPPA>, <https://doi.org/10.1214/20-AOAS1330SUPPB>
- PINAULT, L., TJEPKEMA, M., CROUSE, D. L., WEICHENTHAL, S., VAN DONKELAAR, A., MARTIN, R. V., BRAUER, M., CHEN, H. and BURNETT, R. T. (2016). Risk estimates of mortality attributed to low concentrations of ambient fine particulate matter in the Canadian community health survey cohort. *Environ. Health* **15** Art. ID 18. <https://doi.org/10.1186/s12940-016-0111-6>
- PINAULT, L. L., WEICHENTHAL, S., CROUSE, D. L., BRAUER, M., ERICKSON, A., VAN DONKELAAR, A., MARTIN, R. V., HYSTAD, P., CHEN, H. et al. (2017). Associations between fine particulate matter and mortality in the 2001 Canadian Census Health and Environment Cohort. *Environ. Res.* **159** 406–415. <https://doi.org/10.1016/J.ENVRES.2017.08.037>
- POPE, C. A., CROPPER, M., COGGINS, J. and COHEN, A. (2015). Health benefits of air pollution abatement policy: Role of the shape of the concentration-response function. *J. Air Waste Manage. Assoc.* **65** 516–522. <https://doi.org/10.1080/10962247.2014.993004>
- MACKENZIE ROSS, S. J., BREWIN, C. R., CURRAN, H. V., FURLONG, C. E., ABRAHAM-SMITH, K. M. and HARRISON, V. (2010). Neuropsychological and psychiatric functioning in sheep farmers exposed to low levels of organophosphate pesticides. *Neurotoxicol. Teratol.* **32** 452–459. <https://doi.org/10.1016/j.ntt.2010.03.004>
- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.
- RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.
- SÄVJE, F., ARONOW, P. M. and HUDGENS, M. G. (2018). Average treatment effects in the presence of unknown interference.
- SCHAFER, J. (2015). *causaldrf*: Tools for estimating causal dose response functions.
- SCHOLZE, M., BOEDEKER, W., FAUST, M., BACKHAUS, T., ALTENBURGER, R. and HORST, L. (2001). A general best-fit method for concentration-response curves and the estimation of low-effect concentrations. *Environ. Toxicol. Chem.* **20** 448–457.
- SCHWARTZ, J., BIND, M.-A. and KOUTRAKIS, P. (2017). Estimating causal effects of local air pollution on daily deaths: Effect of low levels. *Environ. Health Perspect.* **125** 23–29. <https://doi.org/10.1289/EHP232>
- SCHWARTZ, J., FONG, K. and ZANOBETTI, A. (2018). A national multicity analysis of the causal effect of local pollution, NO₂, and PM_{2.5} on mortality. *Environ. Health Perspect.* **126** Art. ID 087004.
- SCHWARTZ, J., LADEN, F. and ZANOBETTI, A. (2002). The concentration-response relation between PM_{2.5} and daily deaths. *Environ. Health Perspect.* **110** 1025–1029.
- SHADDICK, G., LEE, D., ZIDEK, J. V. and SALWAY, R. (2008). Estimating exposure response functions using ambient pollution concentrations. *Ann. Appl. Stat.* **2** 1249–1270. [MR2655658 https://doi.org/10.1214/08-AOAS177](https://doi.org/10.1214/08-AOAS177)
- SHI, L., ZANOBETTI, A., KLOOG, I., COULL, B. A., KOUTRAKIS, P., MELLY, S. J. and SCHWARTZ, J. D. (2016). Low-concentration PM_{2.5} and mortality: Estimating acute and chronic effects in a population-based study. *Environ. Health Perspect.* **124** 46–52.
- NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–480.
- THURSTON, G. D., AHN, J., CROMAR, K. R., SHAO, Y., REYNOLDS, H. R., JERRETT, M., LIM, C. C., SHANLEY, R., PARK, Y. et al. (2016). Ambient particulate matter air pollution exposure and mortality in the NIH–AARP diet and health cohort. *Environ. Health Perspect.* **124** 484–490. <https://doi.org/10.1289/ehp.1509676>
- VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** Art. ID 25. [MR2349918 https://doi.org/10.2202/1544-6115.1309](https://doi.org/10.2202/1544-6115.1309)
- VAN DER OOST, R., BEYER, J. and VERMEULEN, N. P. E. (2003). Fish bioaccumulation and biomarkers in environmental risk assessment: A review. *Environ. Toxicol. Pharmacol.* **13** 57–149.

- VANSTEELENDT, S., BEKAERT, M. and CLAESKENS, G. (2012). On model selection and model misspecification in causal inference. *Stat. Methods Med. Res.* **21** 7–30. MR2867536 <https://doi.org/10.1177/0962280210387717>
- VODONOS, A., AWAD, Y. A. and SCHWARTZ, J. (2018). The concentration-response between long-term PM_{2.5} exposure and mortality; a meta-regression approach. *Environ. Res.* **166** 677–689. <https://doi.org/10.1016/j.envres.2018.06.021>
- WANG, C., PARMIGIANI, G. and DOMINICI, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* **68** 661–686. MR3055168 <https://doi.org/10.1111/j.1541-0420.2011.01731.x>
- WANG, C., DOMINICI, F., PARMIGIANI, G. and ZIGLER, C. M. (2015). Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics* **71** 654–665. MR3402601 <https://doi.org/10.1111/biom.12315>
- WANG, Y., SHI, L., LEE, M., LIU, P., DI, Q., ZABONETTI, A. and SCHWARTZ, J. D. (2018). Long-term exposure to PM_{2.5} and mortality among older adults in the southeastern US. *Epidemiology* **28** 207–214. <https://doi.org/10.1097/EDE.0000000000000614> Long-term
- 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
- WEICHENTHAL, S., KULKA, R., LAVIGNE, E., RIJSWIJK, D. V., BRAUER, M., VILLENEUVE, P. J., STIEB, D., JOSEPH, L. and BURNETT, R. T. (2017). Biomass burning as a source of ambient fine particulate air pollution and acute myocardial infarction. *Epidemiology* **28** 329–337. <https://doi.org/10.1097/EDE.0000000000000636>
- WILSON, A. and REICH, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics* **70** 852–861. MR3295746 <https://doi.org/10.1111/biom.12203>
- ZANOBETTI, A. and SCHWARTZ, J. (2007). Particulate air pollution, progression, and survival after myocardial infarction. *Environ. Health Perspect.* **115** 769–775. <https://doi.org/10.1289/ehp.9201>
- ZEGER, S. L., DOMINICI, F., MCDERMOTT, A. and SAMET, J. M. (2008). Mortality in the Medicare population and chronic exposure to fine particulate air pollution in urban centers (2000–2005). *Environ. Health Perspect.* **116** 1614–1619. <https://doi.org/10.1289/ehp.11449>
- ZIGLER, C. M. and DOMINICI, F. (2014). Point: Clarifying policy evidence with potential-outcomes thinking-beyond exposure-response estimation in air pollution epidemiology. *Am. J. Epidemiol.* **180** 1133–1140.

FOCUSED MODEL SELECTION FOR LINEAR MIXED MODELS WITH AN APPLICATION TO WHALE ECOLOGY

BY CÉLINE CUNEN^{1,*}, LARS WALLØE² AND NILS LID HJORT^{1,†}

¹*Department of Mathematics, University of Oslo, *cmlcunen@math.uio.no; †nils@math.uio.no*

²*Institute of Basic Medical Sciences, University of Oslo, lars.walloe@medisin.uio.no*

A central point of disagreement, in certain long-standing discussions about a particular whaling dataset in the Scientific Committee of the International Whaling Commission, has directly involved model selection issues for linear mixed effect models. The biological question under discussion is associated with a clearly defined parameter of primary interest, a focus parameter, which makes model selection with the Focused Information Criterion (FIC) more appropriate than other selection methods. Since the existing FIC methodology has not covered the case of linear mixed effects models, this article sets up the required framework and develops the necessary formulae for the relevant FIC. Our new criterion requires the asymptotic distribution of estimators derived for a given candidate linear mixed model but with behaviour examined under a wider linear mixed model. These results, needed here to build our FIC, also have independent interest.

REFERENCES

- BATES, D., MAECHLER, M., BOLKER, B. and WALKER, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R Package Version 1.1-23.
- BEHL, P., DETTE, H., FRONDEL, M. and TAUCHMANN, H. (2012). Choice is suffering: A focused information criterion for model selection. *Econ. Model.* **29** 817–822.
- BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. [MR3099122 https://doi.org/10.1214/12-AOS1077](https://doi.org/10.1214/12-AOS1077)
- BOLKER, B. M., BROOKS, M. E., CLARK, C. J., GEANGE, S. W., POULSEN, J. R., STEVENS, M. H. H. and WHITE, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends Ecol. Evol.* **24** 127–135.
- BONDELL, H. D., KRISHNA, A. and GHOSH, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66** 1069–1077. [MR2758494 https://doi.org/10.1111/j.1541-0420.2010.01391.x](https://doi.org/10.1111/j.1541-0420.2010.01391.x)
- BROWNLEES, C. T. and GALLO, G. M. (2008). On variable selection for volatility forecasting: The role of focused selection criteria. *J. Financ. Econom.* **6** 513–539.
- CHARKHI, A. and CLAESKENS, G. (2018). Asymptotic post-selection inference for the Akaike information criterion. *Biometrika* **105** 645–664. [MR3842890 https://doi.org/10.1093/biomet/asy018](https://doi.org/10.1093/biomet/asy018)
- CHEN, Z. and DUNSON, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59** 762–769. [MR2025100 https://doi.org/10.1111/j.0006-341X.2003.00089.x](https://doi.org/10.1111/j.0006-341X.2003.00089.x)
- CLAESKENS, G., CROUX, C. and VAN KERCKHOVEN, J. (2007). Prediction-focused model selection for autoregressive models. *Aust. N. Z. J. Stat.* **49** 359–379. [MR2413576 https://doi.org/10.1111/j.1467-842X.2007.00487.x](https://doi.org/10.1111/j.1467-842X.2007.00487.x)
- CLAESKENS, G., CUNEN, C. and HJORT, N. L. (2019). Model selection via focused information criteria for complex data in ecology and evolution. *Front. Ecol. Evol.* **7** 1–13.
- CLAESKENS, G. and HJORT, N. L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.* **98** 900–945. [MR2041482 https://doi.org/10.1198/016214503000000819](https://doi.org/10.1198/016214503000000819)
- CLAESKENS, G. and HJORT, N. L. (2008a). *Model Selection and Model Averaging. Cambridge Series in Statistical and Probabilistic Mathematics* **27**. Cambridge Univ. Press, Cambridge. [MR2431297 https://doi.org/10.1017/CBO9780511790485](https://doi.org/10.1017/CBO9780511790485)
- CLAESKENS, G. and HJORT, N. L. (2008b). Minimizing average risk in regression models. *Econometric Theory* **24** 493–527. [MR2422864 https://doi.org/10.1017/S0266466608080201](https://doi.org/10.1017/S0266466608080201)

- CRAIU, R. V. and DUCHESNE, T. (2018). A scalable and efficient covariate selection criterion for mixed effects regression models with unknown random effects structure. *Comput. Statist. Data Anal.* **117** 154–161. [MR3711487 https://doi.org/10.1016/j.csda.2017.07.011](https://doi.org/10.1016/j.csda.2017.07.011)
- CUNEN, C., WALLØE, L. and HJORT, N. L. (2017). Decline in energy storage in Antarctic minke whales during the JARPA period: assessment via the focused information criterion (FIC). *IWC/SC/67A/EM04*.
- CUNEN, C., WALLØE, L., KONISHI, K. and HJORT, N. L. (2020). Decline in energy storage in the Antarctic minke whale (*Balaenoptera bonaerensis*) in the Southern Ocean during the 1990s. Submitted for publication.
- DEMIDENKO, E. (2013). *Mixed Models: Theory and Applications with R*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. [MR3235905](https://doi.org/10.1002/9781118200542)
- DE LA MARE, W., MCKINLAY, J. and WELSH, A. (2017). Analyses of the JARPA Antarctic minke whale fat weight data set. *IWC/SC/67A/EM01*.
- GANDY, A. and HJORT, N. L. (2013). Focused information criteria for semiparametric linear hazard regression. Technical report, Dept. Mathematics, Univ. Oslo.
- GOVERNMENT OF JAPAN (1987). The program for research on the Southern Hemisphere minke whale and for preliminary research on the marine ecosystem in the Antarctic. *IWC/SC/39/04* [available from the IWC Secretariat].
- GREENLAND, S., PEARL, J., ROBINS, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10** 37–48.
- GRUEBER, C. E., NAKAGAWA, S., LAWS, R. J. and JAMIESON, I. G. (2011). Multimodel inference in ecology and evolution: Challenges and solutions. *J. Evol. Biol.* **24** 699–711.
- GUMEDZE, F. N. and DUNNE, T. T. (2011). Parameter estimation and inference in the linear mixed model. *Linear Algebra Appl.* **435** 1920–1944. [MR2810636 https://doi.org/10.1016/j.laa.2011.04.015](https://doi.org/10.1016/j.laa.2011.04.015)
- HEAGERTY, P. J. and KURLAND, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* **88** 973–985. [MR1872214 https://doi.org/10.1093/biomet/88.4.973](https://doi.org/10.1093/biomet/88.4.973)
- HERMANSEN, G. H., HJORT, N. L. and KJESBU, O. S. (2016). Modern statistical methods applied on extensive historic data: Hjort liver quality time series 1859–2012 and associated influential factors. *Can. J. Fish. Aquat. Sci.* **73** 279–295.
- HJORT, N. L. and CLAESKENS, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98** 879–899. [MR2041481 https://doi.org/10.1198/016214503000000828](https://doi.org/10.1198/016214503000000828)
- HJORT, N. L. and CLAESKENS, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *J. Amer. Statist. Assoc.* **101** 1449–1464. [MR2279471 https://doi.org/10.1198/016214506000000069](https://doi.org/10.1198/016214506000000069)
- HUI, F. K. C., MÜLLER, S. and WELSH, A. H. (2017). Joint selection in mixed models using regularized PQL. *J. Amer. Statist. Assoc.* **112** 1323–1333. [MR3735380 https://doi.org/10.1080/01621459.2016.1215989](https://doi.org/10.1080/01621459.2016.1215989)
- IBRAHIM, J. G., ZHU, H., GARCIA, R. I. and GUO, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67** 495–503. [MR2829018 https://doi.org/10.1111/j.1541-0420.2010.01463.x](https://doi.org/10.1111/j.1541-0420.2010.01463.x)
- JIANG, J., RAO, J. S., GU, Z. and NGUYEN, T. (2008). Fence methods for mixed model selection. *Ann. Statist.* **36** 1669–1692. [MR2435452 https://doi.org/10.1214/07-AOS517](https://doi.org/10.1214/07-AOS517)
- JULLUM, M. and HJORT, N. L. (2017). Parametric or nonparametric: The FIC approach. *Statist. Sinica* **27** 951–981. [MR3699689](https://doi.org/10.1007/s10985-018-9450-7)
- JULLUM, M. and HJORT, N. L. (2019). What price semiparametric Cox regression? *Lifetime Data Anal.* **25** 406–438. [MR3960833 https://doi.org/10.1007/s10985-018-9450-7](https://doi.org/10.1007/s10985-018-9450-7)
- KONISHI, K. and WALLØE, L. (2015). Substantial decline in energy storage and stomach fullness in Antarctic minke whales during the 1990s. *J. Cetacean Res. Manag.* **15** 77–92.
- KONISHI, K., TAMURA, T., ZENITANI, R., BANDO, T., KATO, H. and WALLØE, L. (2008). Decline in energy storage in the Antarctic minke whale (*Balaenoptera bonaerensis*) in the Southern Ocean. *Polar Biol.* **31** 1509–1520.
- LAWS, R. M. (1977). Seals and whales of the Southern Ocean. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **279** 81–96.
- MAGNUS, J. R. and NEUDECKER, H. (1979). The commutation matrix: Some properties and applications. *Ann. Statist.* **7** 381–394. [MR0520247](https://doi.org/10.1007/s10985-018-9450-7)
- MAGNUS, J. R. and NEUDECKER, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, Chichester. [MR0940471](https://doi.org/10.1002/9781118200542)
- MCKINLAY, J., DE LA MARE, W. and WELSH, A. (2017). A re-examination of minke whale body condition as reflected in the data. *IWC/SC/67A/EM02*.
- MORI, M. and BUTTERWORTH, D. S. (2006). A first step towards modelling the krill–predator dynamics of the Antarctic ecosystem. *CCAMLR Sci.* **13** 217–277.
- MÜLLER, S., SCEALY, J. L. and WELSH, A. H. (2013). Model selection in linear mixed models. *Statist. Sci.* **28** 135–167. [MR3112403 https://doi.org/10.1214/12-sts410](https://doi.org/10.1214/12-sts410)

- PENG, H. and LU, Y. (2012). Model selection in linear mixed effect models. *J. Multivariate Anal.* **109** 109–129. [MR2922858 https://doi.org/10.1016/j.jmva.2012.02.005](https://doi.org/10.1016/j.jmva.2012.02.005)
- PINHEIRO, J. and BATES, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, Berlin.
- SCHWEDER, T. and HJORT, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions. Cambridge Series in Statistical and Probabilistic Mathematics* **41**. Cambridge Univ. Press, New York. [MR3558738 https://doi.org/10.1017/CBO9781139046671](https://doi.org/10.1017/CBO9781139046671)
- TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. and TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.* **111** 600–620. [MR3538689 https://doi.org/10.1080/01621459.2015.1108848](https://doi.org/10.1080/01621459.2015.1108848)
- VAIDA, F. and BLANCHARD, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92** 351–370. [MR2201364 https://doi.org/10.1093/biomet/92.2.351](https://doi.org/10.1093/biomet/92.2.351)
- VER HOEF, J. M. and BOVENG, P. L. (2015). Iterating on a single model is a viable alternative to multimodel inference. *J. Wildl. Manag.* **79** 719–729.
- VERBEKE, G. and LESAFFRE, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Comput. Statist. Data Anal.* **23** 541–556. [MR1437679 https://doi.org/10.1016/S0167-9473\(96\)00047-3](https://doi.org/10.1016/S0167-9473(96)00047-3)
- VERBEKE, G., SPIESSENS, B. and LESAFFRE, E. (2001). Conditional linear mixed models. *Amer. Statist.* **55** 25–34. [MR1849040 https://doi.org/10.1198/000313001300339905](https://doi.org/10.1198/000313001300339905)
- ZHANG, X. and LIANG, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *Ann. Statist.* **39** 174–200. [MR2797843 https://doi.org/10.1214/10-AOS832](https://doi.org/10.1214/10-AOS832)

A GLOBAL-LOCAL APPROACH FOR DETECTING HOTSPOTS IN MULTIPLE-RESPONSE REGRESSION

BY HÉLÈNE RUFFIEUX¹, ANTHONY C. DAVISON², JÖRG HAGER³, JAMIE INSHAW⁴,
BENJAMIN P. FAIRFAX⁵, SYLVIA RICHARDSON^{1,6} AND LEONARDO BOTTOLO^{7,6,1}

¹MRC Biostatistics Unit, University of Cambridge, helene.ruffieux@mrc-bsu.cam.ac.uk

²Ecole Polytechnique Fédérale de Lausanne (EPFL), anthony.davison@epfl.ch

³Nestlé Research, EPFL Innovation Park, jorg.hager@rd.nestle.com

⁴Wellcome Centre for Human Genetics, Oxford, University of Oxford, jinshaw@well.ox.ac.uk

⁵Department of Oncology, MRC Weatherall Institute for Molecular Medicine, University of Oxford,
benjamin.fairfax@oncology.ox.ac.uk

⁶Alan Turing Institute, sylvia.richardson@mrc-bsu.cam.ac.uk

⁷Department of Medical Genetics, University of Cambridge, lb664@cam.ac.uk

We tackle modelling and inference for variable selection in regression problems with many predictors and many responses. We focus on detecting *hotspots*, that is, predictors associated with several responses. Such a task is critical in statistical genetics, as hotspot genetic variants shape the architecture of the genome by controlling the expression of many genes and may initiate decisive functional mechanisms underlying disease endpoints. Existing hierarchical regression approaches designed to model hotspots suffer from two limitations: their discrimination of hotspots is sensitive to the choice of top-level scale parameters for the propensity of predictors to be hotspots, and they do not scale to large predictor and response vectors, for example, of dimensions 10^3 – 10^5 in genetic applications. We address these shortcomings by introducing a flexible hierarchical regression framework that is tailored to the detection of hotspots and scalable to the above dimensions. Our proposal implements a fully Bayesian model for hotspots based on the horseshoe shrinkage prior. Its global-local formulation shrinks noise globally and, hence, accommodates the highly sparse nature of genetic analyses while being robust to individual signals, thus leaving the effects of hotspots unshrunk. Inference is carried out using a fast variational algorithm coupled with a novel simulated annealing procedure that allows efficient exploration of multimodal distributions.

REFERENCES

- ALQUIER, P. and RIDGWAY, J. (2017). Concentration of tempered posteriors and of their variational approximations. Preprint. Available at [arXiv:1706.09293](https://arxiv.org/abs/1706.09293).
- ALQUIER, P., RIDGWAY, J. and CHOPIN, N. (2016). On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.* **17** Art. ID 239. [MR3595173](https://arxiv.org/abs/1605.02367)
- BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. [MR2065192](https://doi.org/10.1214/009053604000000238) <https://doi.org/10.1214/009053604000000238>
- BARDENET, R., DOUCET, A. and HOLMES, C. (2014). Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *International Conference on Machine Learning (ICML)* 405–413.
- BERGER, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist.* **8** 716–761. [MR0572619](https://arxiv.org/abs/1706.09293)
- BHADRA, A. and MALLICK, B. K. (2013). Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics* **69** 447–457. [MR3071063](https://doi.org/10.1111/biom.12021) <https://doi.org/10.1111/biom.12021>

Key words and phrases. Annealed variational inference, hierarchical model, horseshoe prior, molecular quantitative trait locus analyses, multiplicity control, normal scale mixture, regulation hotspot, shrinkage, statistical genetics, variable selection.

- BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. (2016). Default Bayesian analysis with global-local shrinkage priors. *Biometrika* **103** 955–969. [MR3620450](#) <https://doi.org/10.1093/biomet/asw041>
- BHATTACHARYA, A. and DUNSON, D. B. (2010). Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika* **97** 851–865. [MR2746156](#) <https://doi.org/10.1093/biomet/asq044>
- 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>
- BOTTOLO, L., PETRETTO, E., BLANKENBERG, S., CAMBIEN, F., COOK, S. A., TIRET, L. and RICHARDSON, S. (2011). Bayesian detection of expression quantitative trait loci hot spots. *Genetics* **189** 1449–1459.
- BRYNEDAL, B., CHOI, J., RAJ, T., BJORNSEN, R., STRANGER, B. E., NEALE, B. M., VOIGHT, B. F. and COTSAPAS, C. (2017). Large-scale trans-eQTLs affect hundreds of transcripts and mediate patterns of transcriptional co-regulation. *Am. J. Hum. Genet.* **100** 581–591. [https://doi.org/10.1016/j.ajhg.2017.02.004](#)
- CARBONETTO, P. and STEPHENS, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.* **7** 73–107. [MR2896713](#) <https://doi.org/10.1214/12-BA703>
- 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>
- DATTA, J. and GHOSH, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Anal.* **8** 111–131. [MR3036256](#) <https://doi.org/10.1214/13-BA805>
- 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.
- FAIRFAX, B. P., HUMBERG, 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** Art. ID 1246949.
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. [MR2221284](#) <https://doi.org/10.1214/06-BA117A>
- GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2** 1360–1383. [MR2655663](#) <https://doi.org/10.1214/08-AOAS191>
- GEORGE, E. I. (2000). The variable selection problem. *J. Amer. Statist. Assoc.* **95** 1304–1308. [MR1825282](#) <https://doi.org/10.2307/2669776>
- GHOSH, P., TANG, X., GHOSH, M. and CHAKRABARTI, A. (2016). Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. *Bayesian Anal.* **11** 753–796. [MR3498045](#) <https://doi.org/10.1214/15-BA973>
- GILAD, Y., RIFKIN, S. A. and PRITCHARD, J. K. (2008). Revealing the architecture of gene regulation: The promise of eQTL studies. *Trends Genet.* **24** 408–415.
- GRAMACY, R., SAMWORTH, R. and KING, R. (2010). Importance tempering. *Stat. Comput.* **20** 1–7. [MR2578072](#) <https://doi.org/10.1007/s11222-008-9108-5>
- GUHANIYOGI, R., QAMAR, S. and DUNSON, D. B. (2018). Bayesian conditional density filtering. *J. Comput. Graph. Statist.* **27** 657–672. [MR3863766](#) <https://doi.org/10.1080/10618600.2017.1422431>
- JIA, Z. and XU, S. (2007). Mapping quantitative trait loci for expression abundance. *Genetics* **176** 611–623.
- KATAHIRA, K., WATANABE, K. and OKADA, M. (2008). Deterministic annealing variant of variational Bayes method. *J. Phys., Conf. Ser.* **95** Art. ID 012015.
- 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** Art. ID 5236.
- 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>
- 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** Art. ID 1246980.
- LEWIN, A., SAADI, H., PETERS, J. E., MORENO-MORAL, A., LEE, J. C., SMITH, K. G. C., PETRETTO, E., BOTTOLO, L. and RICHARDSON, S. (2015). MT-HESS: An efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues. *Bioinformatics* **32** 523–532.
- LI, H. and PATI, D. (2017). Variable selection using shrinkage priors. *Comput. Statist. Data Anal.* **107** 107–119. [MR3575062](#) <https://doi.org/10.1016/j.csda.2016.10.008>
- MACKAY, T. F. C., STONE, E. A. and AYROLES, J. F. (2009). The genetics of quantitative traits: Challenges and prospects. *Nat. Rev. Genet.* **10** 565–577.

- MANDT, S., MCINERNEY, J., ABROL, F., RANGANATH, R. and BLEI, D. (2016). Variational tempering. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research (PMLR)* **51** 704–712.
- 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.
- MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. [MR0997578](#)
- NEVILLE, S. E., ORMEROD, J. T. and WAND, M. P. (2014). Mean field variational Bayes for continuous sparse signal shrinkage: Pitfalls and remedies. *Electron. J. Stat.* **8** 1113–1151. [MR3263115](#) <https://doi.org/10.1214/14-EJS910>
- NICA, A. C. and DERMITZAKIS, E. T. (2013). Expression quantitative trait loci: Present and future. *Philos. Trans. R. Soc. B* **368** Art. ID 20120362.
- O'BRIEN, S. M. and DUNSON, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics* **60** 739–746. [MR2089450](#) <https://doi.org/10.1111/j.0006-341X.2004.00224.x>
- OPPER, M. and SAAD, D., eds. (2001). *Advanced Mean Field Methods: Theory and Practice. Neural Information Processing Series*. MIT Press, Cambridge, MA. [MR1863214](#)
- PARK, J.-H., GAIL, M. H., WEINBERG, C. R., CARROLL, R. J., CHUNG, C. C., WANG, Z., CHANOCK, S. J., FRAUMENI, J. F. and CHATTERJEE, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc. Natl. Acad. Sci. USA* **108** 18026–18031.
- PETERSON, C. B., BOGOMOLOV, M., BENJAMINI, Y. and SABATTI, C. (2016). TreeQTL: Hierarchical error control for eQTL findings. *Bioinformatics* **32** 2556–2558.
- PETRETTO, E., BOTTOLO, L., LANGLEY, S. R., HEINIG, M., MCDERMOTT-ROE, C., SARWAR, R., PRAVENEC, M., HÜBNER, N., AITMAN, T. J. et al. (2010). New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Comput. Biol.* **6** Art. ID e1000737.
- 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. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics 9* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 501–538. Oxford Univ. Press, Oxford. [MR3204017](#) <https://doi.org/10.1093/acprof:oso/9780199694587.003.0017>
- RICHARDSON, S., BOTTOLO, L. and ROSENTHAL, J. S. (2011). Bayesian models for sparse regression analysis of high dimensional data. In *Bayesian Statistics 9* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 539–568. Oxford Univ. Press, Oxford. [MR3204018](#) <https://doi.org/10.1093/acprof:oso/9780199694587.003.0018>
- 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>
- ROSE, K., GUREWITZ, E. and FOX, G. (1990). A deterministic annealing approach to clustering. *Pattern Recogn. Lett.* **11** 589–594.
- ROTIVAL, M., ZELLER, T., WILD, P. S., MAUCHE, S., SZYMCAK, S., SCHILLERT, A., CASTAGNÉ, R., DEISEROTH, A., PROUST, C. et al. (2011). Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. *PLoS Genet.* **7** Art. ID e1002367.
- RUFFIEUX, H., DAVISON, A. C., HAGER, J. and IRINCHEEVA, I. (2017). Efficient inference for genetic association studies with multiple outcomes. *Biostatistics* **18** 618–636. [MR3984082](#) <https://doi.org/10.1093/biostatistics/kxx007>
- RUFFIEUX, H., DAVISON, A. C., HAGER, J., INSHAW, J., FAIRFAX, B., RICHARDSON, S. and BOTTOLO, L. (2020). Supplement to “A global-local approach for detecting hotspots in multiple-response regression”. <https://doi.org/10.1214/20-AOAS1332SUPPA>, <https://doi.org/10.1214/20-AOAS1332SUPPB>
- 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>
- SHABALIN, A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28** 1353–1358.
- SOLOVIEFF, N., COTSAPAS, C., LEE, P. H., PURCELL, S. M. and SMOLLER, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nat. Rev. Genet.* **14** 483–495.
- STRAWDERMAN, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Stat.* **42** 385–388. [MR0397939](#) <https://doi.org/10.1214/aoms/1177693528>
- TAK, Y. G. and FARNHAM, P. J. (2015). Making sense of GWAS: Using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenet. Chromatin* **8** Art. ID 57. <https://doi.org/10.1186/s13072-015-0050-4>
- UEDA, N. and NAKANO, R. (1998). Deterministic annealing EM algorithm. *Neural Netw.* **11** 271–282.

- 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>
- VAN DER PAS, S. L., SALOMOND, J.-B. and SCHMIDT-HIEBER, J. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electron. J. Stat.* **10** 976–1000. [MR3486423](#) <https://doi.org/10.1214/16-EJS1130>
- VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Adaptive posterior contraction rates for the horseshoe. *Electron. J. Stat.* **11** 3196–3225. [MR3705450](#) <https://doi.org/10.1214/17-EJS1316>
- VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2016). How many needles in the haystack? Adaptive inference and uncertainty quantification for the horseshoe. Preprint. Available at [arXiv:1607.01892](#).
- WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.
- WANG, X. and DUNSON, D. B. (2013). Parallelizing MCMC via Weierstrass sampler. Preprint. Available at [arXiv:1312.4605](#).
- WARD, L. D. and KELLIS, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30** 1095–1106. <https://doi.org/10.1038/nbt.2422>
- WESTRA, H.-J., PETERS, M. J., ESKO, T., YAGHOOTKAR, H., SCHURMANN, C., KETTUNEN, J., CHRISTIANSEN, M. W., FAIRFAX, B. P., SCHRAMM, K. et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45** 1238–1243.
- YANG, Y., PATI, D. and BHATTACHARYA, A. (2017). Alpha-variational inference with statistical guarantees. Preprint. Available at [arXiv:1710.03266](#).
- 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.
- YIN, J. and LI, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5** 2630–2650. [MR2907129](#) <https://doi.org/10.1214/11-AOAS494>

SEASONAL WARRANTY PREDICTION BASED ON RECURRENT EVENT DATA

BY QIANQIAN SHAN^{1,*}, YILI HONG² AND WILLIAM Q. MEEKER^{1,†}

¹Department of Statistics, Iowa State University, *qshan@iastate.edu; †wqmeeker@iastate.edu

²Department of Statistics, Virginia Tech, yilihong@vt.edu

Warranty return data from repairable systems, such as home appliances, lawn mowers, computers and automobiles, result in recurrent event data. The nonhomogeneous Poisson process (NHPP) model is used widely to describe such data. Seasonality in the repair frequencies and other variabilities, however, complicate the modeling of recurrent event data. Not much work has been done to address the seasonality, and this paper provides a general approach for the application of NHPP models with dynamic covariates to predict seasonal warranty returns. The methods presented here, however, can be applied to other applications that result in seasonal recurrent event data. A hierarchical clustering method is used to stratify the population into groups that are more homogeneous than the overall population. The stratification facilitates modeling the recurrent event data with both time-varying and time-constant covariates. We demonstrate and validate the models using warranty claims data for two different types of products. The results show that our approach provides important improvements in the predictive power of monthly events compared with models that do not take the seasonality and covariates into account.

REFERENCES

- AALEN, O. O. (1988). Heterogeneity in survival analysis. *Stat. Med.* **7** 1121–1137.
- AHN, C.-W., CHAE, K.-C. and CLARK, G. M. (1998). Estimating parameters of the power law process with two measures of failure time. *J. Qual. Technol.* **30** 127–132.
- BERAN, R. (1990). Calibrating prediction regions. *J. Amer. Statist. Assoc.* **85** 715–723. [MR1138352](#)
- BRÉMAUD, P. (1981). *Point Processes and Queues: Martingale Dynamics*. Springer Series in Statistics. Springer, New York–Berlin. [MR0636252](#)
- CIFUENTES-AMADO, M. V. and CEPEDA-CUERVO, E. (2015). Non-homogeneous Poisson process to model seasonal events: Application to the health diseases. *Int. J. Stat. Med. Res.* **4** 337–346.
- COOK, R. J. and LAWLESS, J. F. (2007). *The Statistical Analysis of Recurrent Events*. Statistics for Biology and Health. Springer, New York. [MR3822124](#)
- DUCHESNE, T. and LAWLESS, J. (2000). Alternative time scales and failure time models. *Lifetime Data Anal.* **6** 157–179. [MR1766199](#) <https://doi.org/10.1023/A:1009616111968>
- FONSECA, G., GIUMMOLÈ, F. and VIDONI, P. (2014). Calibrating predictive distributions. *J. Stat. Comput. Simul.* **84** 373–383. [MR3169333](#) <https://doi.org/10.1080/00949655.2012.709518>
- FREDETTE, M. and LAWLESS, J. F. (2007). Finite-horizon prediction of recurrent events, with application to forecasts of warranty claims. *Technometrics* **49** 66–80. [MR2345453](#) <https://doi.org/10.1198/004017006000000390>
- GRANDELL, J. (1997). *Mixed Poisson Processes*. Monographs on Statistics and Applied Probability **77**. CRC Press, London. [MR1463943](#) <https://doi.org/10.1007/978-1-4899-3117-7>
- HAMADA, M. S., WILSON, A. G., REESE, C. S. and MARTZ, H. F. (2008). *Bayesian Reliability*. Springer Series in Statistics. Springer, New York. [MR2724437](#) <https://doi.org/10.1007/978-0-387-77950-8>
- HARTIGAN, J. A. (1975). *Clustering Algorithms*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York. [MR0405726](#)
- HARTIGAN, J. A. and WONG, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **28** 100–108.

Key words and phrases. EM algorithm, hierarchical clustering, missing data, NHPP, random effects, seasonal dynamic covariates.

- 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](https://doi.org/10.1007/978-0-387-84858-7)
- JAMES, G., WITTEN, D., HASTIE, T. and TIBSHIRANI, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. *Springer Texts in Statistics* **103**. Springer, New York. [MR3100153 https://doi.org/10.1007/978-1-4614-7138-7](https://doi.org/10.1007/978-1-4614-7138-7)
- KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481. [MR0093867](https://doi.org/10.1080/01621459.1958.10501452)
- KARYAGINA, M., WONG, W. and VLACIC, L. (1998). Life cycle cost modeling using marked point processes. *Reliab. Eng. Syst. Saf.* **59** 291–298.
- KAUFMAN, L. and ROUSSEEUW, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. [MR1044997 https://doi.org/10.1002/9780470316801](https://doi.org/10.1002/9780470316801)
- KLEIN, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48** 795–806.
- KOUTSELLIS, T., MOURELATOS, Z., HIJAWI, M., GUO, H. and CASTANIER, M. (2017). Warranty forecasting of repairable systems for different production patterns. *SAE Int. J. Mater. Manuf.* **10** 264–273.
- KRIVTSOV, V. V. (2007). Practical extensions to NHPP application in repairable system reliability analysis. *Reliab. Eng. Syst. Saf.* **92** 560–562.
- LAWLESS, J. F. (1987). Regression methods for Poisson process data. *J. Amer. Statist. Assoc.* **82** 808–815. [MR0909986](https://doi.org/10.1080/01621459.1987.10477444)
- LAWLESS, J. F. and CROWDER, M. J. (2010). Models and estimation for systems with recurrent events and usage processes. *Lifetime Data Anal.* **16** 547–570. [MR2726224 https://doi.org/10.1007/s10985-010-9157-x](https://doi.org/10.1007/s10985-010-9157-x)
- LAWLESS, J. F. and FREDETTE, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika* **92** 529–542. [MR2202644 https://doi.org/10.1093/biomet/92.3.529](https://doi.org/10.1093/biomet/92.3.529)
- LAWLESS, J. F. and NADEAU, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37** 158–168. [MR1333194 https://doi.org/10.2307/1269617](https://doi.org/10.2307/1269617)
- LLOYD, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28** 129–137. [MR0651807 https://doi.org/10.1109/TIT.1982.1056489](https://doi.org/10.1109/TIT.1982.1056489)
- MEEKER, W. Q. and ESCOBAR, L. A. (1998). *Statistical Methods for Reliability Data*. Wiley, New York.
- NELSON, W. (1988). Analysis of repair data. In 1988. *Proceedings., Annual Reliability and Maintainability Symposium, IEEE*. <https://doi.org/10.1109/ARMS.1988.196451>.
- NELSON, W. B. (2003). *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*. *ASA-SIAM Series on Statistics and Applied Probability* **10**. SIAM, Philadelphia, PA. [MR1953176 https://doi.org/10.1137/1.9780898718454](https://doi.org/10.1137/1.9780898718454)
- NGAILO, T., SHABAN, N., REUDER, J., RUTALEBWA, E. and MUGUME, I. (2016). Non homogeneous Poisson process modelling of seasonal extreme rainfall events in Tanzania. *Int. J. Sci. Res.* **5** 1858–1868.
- RAI, B. K. (2009). Warranty spend forecasting for subsystem failures influenced by calendar month seasonality. *IEEE Trans. Reliab.* **58** 649–657.
- RIGDON, S. E. and BASU, A. P. (2000). *Statistical Methods for the Reliability of Repairable Systems*. *Wiley Series in Probability and Statistics: Applied Probability and Statistics*. Wiley, New York. [MR1781523](https://doi.org/10.1002/9781118133333)
- ROSS, S. M. (2014). *Introduction to Probability Models*, 11th ed. Elsevier/Academic Press, Amsterdam. [MR3307945](https://doi.org/10.1016/B978-0-12-375024-2)
- RYAN, K. J., HAMADA, M. S. and REESE, C. S. (2011). A Bayesian hierarchical power law process model for multiple repairable systems with an application to supercomputer reliability. *J. Qual. Technol.* **43** 209–223.
- SHAN, Q., HONG, Y. and MEEKER, W. Q. (2020). Supplement to “Seasonal warranty prediction based on recurrent event data.” <https://doi.org/10.1214/20-AOAS1333SUPPA>, <https://doi.org/10.1214/20-AOAS1333SUPPB>
- SLIMACEK, V. and LINDQVIST, B. (2016). Reliability of wind turbines modeled by a Poisson process with covariates, unobserved heterogeneity and seasonality. *Wind Energy* **19** 1991–2002.
- TEERAPABOLARN, K. (2014). Poisson approximation for independent negative binomial random variables. *Int. J. Pure Appl. Math.* **93** 779–781.
- THERNEAU, T. M., GRAMBSCH, P. M. and PANKRATZ, V. S. (2003). Penalized survival models and frailty. *J. Comput. Graph. Statist.* **12** 156–175. [MR1965213 https://doi.org/10.1198/1061860031365](https://doi.org/10.1198/1061860031365)
- WU, S. (2012). Warranty data analysis: A review. *Qual. Reliab. Eng. Int.* **28** 795–805.
- XIAO, S., KOTTAS, A. and SANSÓ, B. (2015). Modeling for seasonal marked point processes: An analysis of evolving hurricane occurrences. *Ann. Appl. Stat.* **9** 353–382. [MR3341119 https://doi.org/10.1214/14-AOAS796](https://doi.org/10.1214/14-AOAS796)
- XU, Z., HONG, Y. and MEEKER, W. Q. (2015). Assessing risk of a serious failure mode based on limited field data. *IEEE Trans. Reliab.* **64** 51–62.

GENERALIZED ACCELERATED RECURRENCE TIME MODEL IN THE PRESENCE OF A DEPENDENT TERMINAL EVENT

BY BO WEI^{1,*}, ZHUMIN ZHANG^{2,‡}, HUICHUAN J. LAI^{2,§} AND LIMIN PENG^{1,†}

¹Department of Biostatistics and Bioinformatics, Emory University, *bo.wei@emory.edu; †lpeng@emory.edu

²Departments of Nutritional Sciences, University of Wisconsin–Madison, ‡zhuminzhang@wisc.edu; §hlai@wisc.edu

Recurrent events are commonly encountered in longitudinal studies. The observation of recurrent events is often stopped by a dependent terminal event in practice. For this data scenario, we propose two sensible adaptations of the generalized accelerated recurrence time (GART) model (*J. Amer. Statist. Assoc.* **111** (2016) 145–156) to provide useful alternative analyses that can offer physical interpretations while rendering extra flexibility beyond the existing work based on the accelerated failure time model. Our modeling strategies align with the rationale underlying the use of the survivors' rate function or the adjusted rate function to account for the presence of the dependent terminal event. For the proposed models, we identify and develop estimation and inference procedures which can be readily implemented based on existing software. We establish the asymptotic properties of the new estimator. Simulation studies demonstrate good finite-sample performance of the proposed methods. An application to a dataset from the Cystic Fibrosis Foundation Patient Registry (CFFPR) illustrates the practical utility of the new methods.

REFERENCES

- ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120. [MR0673646](#)
- COOK, R. J. and LAWLESS, J. F. (1997). Marginal analysis of recurrent events and a terminating event. *Stat. Med.* **16** 911–924.
- DAVIES, J. C. (2002). *Pseudomonas aeruginosa* in cystic fibrosis: Pathogenesis and persistence. *Paediatr. Respir. Rev.* **3** 128–134.
- FYGENSON, M. and RITOV, Y. (1994). Monotone estimating equations for censored data. *Ann. Statist.* **22** 732–746. [MR1292538](#) <https://doi.org/10.1214/aos/1176325493>
- GHOSH, D. and LIN, D. Y. (2002). Marginal regression models for recurrent and terminal events. *Statist. Sinica* **12** 663–688. [MR1929958](#)
- GHOSH, D. and LIN, D. Y. (2003). Semiparametric analysis of recurrent events data in the presence of dependent censoring. *Biometrics* **59** 877–885. [MR2025111](#) <https://doi.org/10.1111/j.0006-341X.2003.00102.x>
- HELTSHE, S. L., KHAN, U., BECKETT, V., BAINES, A., EMERSON, J., SANDERS, D. B., GIBSON, R. L., MORGAN, W. and ROSENFELD, M. (2018). Longitudinal development of initial, chronic and mucoid *Pseudomonas aeruginosa* infection in young children with cystic fibrosis. *J. Cyst. Fibros.* **17** 341–347. <https://doi.org/10.1016/j.jcf.2017.10.008>
- HUANG, X. and LIU, L. (2007). A joint frailty model for survival and gap times between recurrent events. *Biometrics* **63** 389–397. [MR2370797](#) <https://doi.org/10.1111/j.1541-0420.2006.00719.x>
- HUANG, Y. and PENG, L. (2009). Accelerated recurrence time models. *Scand. J. Stat.* **36** 636–648. [MR2573300](#) <https://doi.org/10.1111/j.1467-9469.2009.00645.x>
- HUANG, C.-Y. and WANG, M.-C. (2004). Joint modeling and estimation for recurrent event processes and failure time data. *J. Amer. Statist. Assoc.* **99** 1153–1165. [MR2109503](#) <https://doi.org/10.1198/016214504000001033>
- 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>
- KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. Wiley Series in Probability and Statistics. Wiley Interscience, Hoboken, NJ. [MR1924807](#) <https://doi.org/10.1002/9781118032985>

- LI, R. and PENG, L. (2011). Quantile regression for left-truncated semicompeting risks data. *Biometrics* **67** 701–710. [MR2829124 https://doi.org/10.1111/j.1541-0420.2010.01521.x](https://doi.org/10.1111/j.1541-0420.2010.01521.x)
- LIN, D. Y., SUN, W. and YING, Z. (1999). Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika* **86** 59–70. [MR1688071 https://doi.org/10.1093/biomet/86.1.59](https://doi.org/10.1093/biomet/86.1.59)
- LIN, D. Y., WEI, L. J. and YING, Z. (1998). Accelerated failure time models for counting processes. *Biometrika* **85** 605–618. [MR1665810 https://doi.org/10.1093/biomet/85.3.605](https://doi.org/10.1093/biomet/85.3.605)
- LIN, D. Y., WEI, L. J., YANG, I. and YING, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 711–730. [MR1796287 https://doi.org/10.1111/1467-9868.00259](https://doi.org/10.1111/1467-9868.00259)
- LIU, L., WOLFE, R. A. and HUANG, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics* **60** 747–756. [MR2089451 https://doi.org/10.1111/j.0006-341X.2004.00225.x](https://doi.org/10.1111/j.0006-341X.2004.00225.x)
- LUO, X., WANG, M.-C. and HUANG, C.-Y. (2010). A comparison of various rate functions of a recurrent event process in the presence of a terminal event. *Stat. Methods Med. Res.* **19** 167–182. [MR2649719 https://doi.org/10.1177/0962280208090220](https://doi.org/10.1177/0962280208090220)
- LYCZAK, J. B., CANNON, C. L. and PIER, G. B. (2002). Lung infections associated with cystic fibrosis. *Clin. Microbiol. Rev.* **15** 194–222. <https://doi.org/10.1128/cmr.15.2.194-222.2002>
- MATHEE, K., CIOFU, O., STERNBERG, C., LINDUM, P. W., CAMPBELL, J. I., JENSEN, P., JOHNSEN, A. H., GIVSKOV, M., OHMAN, D. E. et al. (1999). Mucoid conversion of *Pseudomonas aeruginosa* by hydrogen peroxide: A mechanism for virulence activation in the cystic fibrosis lung. *Microbiology* **145** 1349–1357.
- MILOSLAVSKY, M., KELEŞ, S., VAN DER LAAN, M. J. and BUTLER, S. (2004). Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 239–257. [MR2035769 https://doi.org/10.1111/j.1467-9868.2004.00442.x](https://doi.org/10.1111/j.1467-9868.2004.00442.x)
- OLIVEIRA, M. C. L. A., REIS, F. J. C., MONTEIRO, A. P. A. F. and PENNA, F. J. (2002). Effect of meconium ileus on the clinical prognosis of patients with cystic fibrosis. *Braz. J. Med. Biol. Res.* **35** 31–38.
- PENG, L. and FINE, J. P. (2009). Competing risks quantile regression. *J. Amer. Statist. Assoc.* **104** 1440–1453. [MR2596999 https://doi.org/10.1198/jasa.2009.tm08228](https://doi.org/10.1198/jasa.2009.tm08228)
- PEPE, M. S. (1991). Inference for events with dependent risks in multiple endpoint studies. *J. Amer. Statist. Assoc.* **86** 770–778. [MR1147104](https://doi.org/10.1111/j.1541-0420.1991.tb04104.x)
- PEPE, M. S. and CAI, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *J. Amer. Statist. Assoc.* **88** 811–820.
- PRENTICE, R. L., WILLIAMS, B. J. and PETERSON, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika* **68** 373–379. [MR0626396 https://doi.org/10.1093/biomet/68.2.373](https://doi.org/10.1093/biomet/68.2.373)
- PRENTICE, R. L., KALBFLEISCH, J. D., PETERSON JR, A. V., FLOURNOY, N., FAREWELL, V. T. and BRESLOW, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* 541–554.
- REID, N. (1994). A conversation with Sir David Cox. *Statist. Sci.* **9** 439–455. [MR1325436](https://doi.org/10.1111/j.1541-0420.1994.tb00436.x)
- ROBINS, J. M. and ROTNITZKY, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology* 297–331. Springer, Berlin.
- RUSSELL, P. J., HERTZ, P. E. and MCMILLAN, B. (2012). *Biology: The Dynamic Science*. Cengage Learning, Boston, MA.
- SAWYER, S. M., TAYLOR, R., MACMAHON, R. and ROBERTSON, C. F. (1994). Meconium ileus in cystic fibrosis. *Pediatr. Surg. Int.* **9** 180–184.
- SCHAUBEL, D. E. and CAI, J. (2005). Analysis of clustered recurrent event data with application to hospitalization rates among renal failure patients. *Biostatistics* **6** 404–419.
- SUN, X., PENG, L., HUANG, Y. and LAI, H. J. (2016). Generalizing quantile regression for counting processes with applications to recurrent events. *J. Amer. Statist. Assoc.* **111** 145–156. [MR3494649 https://doi.org/10.1080/01621459.2014.995795](https://doi.org/10.1080/01621459.2014.995795)
- VAN DER VAART, A. W. and WELLNER, J. A. (1997). Weak convergence and empirical processes with applications to statistics. *J. Roy. Statist. Soc. Ser. A* **160** 596–608.
- WEI, B., ZHANG, Z., LAI, H. and PENG, L. (2020). Supplement to “Generalized accelerated recurrence time model in the presence of a dependent terminal event.” <https://doi.org/10.1214/20-AOAS1335SUPP>
- YE, Y., KALBFLEISCH, J. D. and SCHAUBEL, D. E. (2007). Semiparametric analysis of correlated recurrent and terminal events. *Biometrics* **63** 78–87, 311. [MR2345577 https://doi.org/10.1111/j.1541-0420.2006.00677.x](https://doi.org/10.1111/j.1541-0420.2006.00677.x)
- ZENG, D. and CAI, J. (2010). A semiparametric additive rate model for recurrent events with an informative terminal event. *Biometrika* **97** 699–712. [MR2672492 https://doi.org/10.1093/biomet/asq039](https://doi.org/10.1093/biomet/asq039)

EARLY IDENTIFICATION OF AN IMPENDING ROCKSLIDE LOCATION VIA A SPATIALLY-AIDED GAUSSIAN MIXTURE MODEL

BY SHUO ZHOU^{1,*}, HOWARD BONDELL^{1,†}, ANTOINETTE TORDESILLAS^{1,‡},
BENJAMIN I. P. RUBINSTEIN^{2,§} AND JAMES BAILEY^{2,¶}

¹*School of Mathematics and Statistics, The University of Melbourne, *shuo.zhou@unimelb.edu.au;*

†howard.bondell@unimelb.edu.au; ‡atordes@unimelb.edu.au

²*School of Computing and Information Systems, The University of Melbourne, §brubinstein@unimelb.edu.au;*

¶baileyj@unimelb.edu.au

Movement of soil and rocks in an unstable slope under gravitational forces is an example of a complex system that is highly dynamic in space and time. A typical failure in such systems is a landslide. Fundamental studies of granular media failure combined with a complex network analysis of radar monitoring data show that distinct partitions emerge in the kinematic field in the early stages of the prefailure regime, and these patterns yield clues to the ultimate location of failure. In this study we address this partitioning of constituent units in complex systems by clustering the kinematic data, specifically, with a Gaussian mixture model. In addition, we assume that neighboring units should move together. As a result, spatial information is taken into account in our model so that spatial proximity is retained. Our case study of a rockslide from high resolution radar monitoring data shows that, by incorporating spatial information, our approach is more effective in revealing the dynamics of the system and detecting the location of a potential landslide, compared to the use of only the kinematics.

REFERENCES

- BRENNING, A. (2005). Spatial prediction models for landslide hazards: Review, comparison and evaluation. *Nat. Hazards Earth Syst. Sci.* **5** 853–862.
- BUI, D. T., PRADHAN, B., LOFMAN, O., REVHAUG, I. and DICK, O. B. (2012). Landslide susceptibility mapping at Hoa Binh province (Vietnam) using an adaptive neuro-fuzzy inference system and GIS. *Comput. Geosci.* **45** 199–211.
- CASAGLI, N., CATANI, F., DEL VENTISETTE, C. and LUZI, G. (2010). Monitoring, prediction, and early warning using ground-based radar interferometry. *Landslides* **7** 291–301.
- CHEN, H. and ZENG, Z. (2013). Deformation prediction of landslide based on improved back-propagation neural network. *Cogn. Comput.* **5** 56–62.
- CHEN, W., XIE, X., WANG, J., PRADHAN, B., HONG, H., BUI, D. T., DUAN, Z. and MA, J. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena* **151** 147–160.
- 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](https://doi.org/10.2307/2344682)
- DI TRAGLIA, F. D., BATTAGLIA, M., NOLESINI, T., LAGOMARSINO, D. and CASAGLI, N. (2015). Shifts in the eruptive styles at Stromboli in 2010–2014 revealed by ground-based InSAR data. *Sci. Rep.* **5** Art. ID 13569. <https://doi.org/10.1038/srep13569>
- DI TRAGLIA, F., CALVARI, S., D’AURIA, L., NOLESINI, T., BONACCORSO, A., FORNACIAI, A., ESPOSITO, A., CRISTALDI, A., FAVALLI, M. and CASAGLI, N. (2018). The 2014 effusive eruption at Stromboli: New insights from in situ and remote-sensing measurements. *Remote Sens.* **10** Art. ID 2035.
- DICK, G. J., EBERHARDT, E., CABREJO-LIÉVANO, A. G., STEAD, D. and ROSE, N. D. (2014). Development of an early-warning time-of-failure analysis methodology for open-pit mine slopes utilizing ground-based slope stability radar monitoring data. *Can. Geotech. J.* **52** 515–529.
- 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](https://doi.org/10.1198/016214502760047131)

- GEBRU, I. D., ALAMEDA-PINEDA, X., FORBES, F. and HORAUD, R. (2016). EM algorithms for weighted-data clustering with application to audio-visual scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **38** 2402–2415. <https://doi.org/10.1109/TPAMI.2016.2522425>
- GOETZ, J. N., BRENNING, A., PETSCHKO, H. and LEOPOLD, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Comput. Geosci.* **81** 1–11.
- HONG, H., POURGHASEMI, H. R. and POURTAGHI, Z. S. (2016). Landslide susceptibility assessment in Lianhua County (China): A comparison between a random forest data mining technique and bivariate and multivariate statistical models. *Geomorphology* **259** 105–118.
- KARIMI, K. and BARRAT, J.-L. (2018). Correlation and shear bands in a plastically deformed granular medium. *Sci. Rep.* **8** Art. ID 4021.
- LAFAYETTE, L. and WIEBELT, B. (2017). Spartan and nemo: Two HPC-cloud hybrid implementations. In 2017 *IEEE 13th International Conference on E-Science (e-Science)* 458–459. IEEE, Piscataway, NJ.
- LE BOUIL, A., AMON, A., SANGLEBOEUF, J.-C., ORAIN, H., BÉSUELLE, P., VIGGIANI, G., CHASLE, P. and CRASSOUS, J. (2014). A biaxial apparatus for the study of heterogeneous and intermittent strains in granular materials. *Granul. Matter* **16** 1–8.
- MIAO, F., WU, Y., XIE, Y. and LI, Y. (2018). Prediction of landslide displacement with step-like behavior based on multialgorithm optimization and a support vector regression model. *Landslides* **15** 475–488.
- NEAUPANE, K. M. and ACHET, S. H. (2004). Use of backpropagation neural network for landslide monitoring: A case study in the higher Himalaya. *Eng. Geol.* **74** 213–226.
- NEWTON, M. A. and RAFTERY, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. Roy. Statist. Soc. Ser. B* **56** 3–48. [MR1257793](https://doi.org/10.2307/2346178)
- O'HAGAN, A., MURPHY, T. B., SCRUGA, L. and GORMLEY, I. C. (2019). Investigation of parameter uncertainty in clustering using a Gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap. *Comput. Statist.* **34** 1779–1813. [MR4030442 https://doi.org/10.1007/s00180-019-00897-9](https://doi.org/10.1007/s00180-019-00897-9)
- PHAM, B. T., BUI, D. T., POURGHASEMI, H. R., INDRA, P. and DHOLAKIA, M. B. (2017). Landslide susceptibility assessment in the Uttarakhand area (India) using GIS: A comparison study of prediction capability of naïve Bayes, multilayer perceptron neural networks, and functional trees methods. *Theor. Appl. Climatol.* **128** 255–273.
- PIENAAR, A. and ANTON, J. (2013). Guidelines for deriving alarm settings based on pre-determined criteria using the movement and surveying radar. *Aust. Cent. Geomech. Newsl.* **41** 7–12.
- POURGHASEMI, H. R. and RAHMATI, O. (2018). Prediction of the landslide susceptibility: Which algorithm, which precision? *Catena* **162** 177–192.
- REICH, B. J. and BONDELL, H. D. (2011). A spatial Dirichlet process mixture model for clustering population genetics data. *Biometrics* **67** 381–390. [MR2829007 https://doi.org/10.1111/j.1541-0420.2010.01484.x](https://doi.org/10.1111/j.1541-0420.2010.01484.x)
- REICHENBACH, P., ROSSI, M., MALAMUD, B. D., MIHIR, M. and GUZZETTI, F. (2018). A review of statistically-based landslide susceptibility models. *Earth-Sci. Rev.* **180** 60–91.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. [MR0468014](https://doi.org/10.2307/2346178)
- THANH, L. N. and DE SMEDT, F. (2014). Slope stability analysis using a physically based model: A case study from A Luoi district in Thua Thien-Hue Province, Vietnam. *Landslides* **11** 897–907.
- TORDESILLAS, A., ZHOU, Z. and BATTERHAM, R. (2018). A data-driven complex systems approach to early prediction of landslides. *Mech. Res. Commun.* **92** 137–141.
- TORDESILLAS, A., WALKER, D. M., ANDÒ, E. and VIGGIANI, G. (2013). Revisiting localized deformation in sand with complex systems. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **469** Art. ID 20120606.
- 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](https://doi.org/10.26434/chemrxiv-2010-01484)
- VITONE, C., COTECCHIA, F., VIGGIANI, G. and HALL, S. A. (2013). Strain fields and mechanical response of a highly to medium fissured bentonite clay. *Int. J. Numer. Anal. Methods Geomech.* **37** 1510–1534.
- WASOWSKI, J. and BOVENGA, F. (2014). Investigating landslides and unstable slopes with satellite multi temporal interferometry: Current issues and future perspectives. *Eng. Geol.* **174** 103–138.
- YAO, X., THAM, L. G. and DAI, F. C. (2008). Landslide susceptibility mapping based on support vector machine: A case study on natural slopes of Hong Kong, China. *Geomorphology* **101** 572–582.
- YAO, W., ZENG, Z., LIAN, C. and TANG, H. (2015). Training enhanced reservoir computing predictor for landslide displacement. *Eng. Geol.* **188** 101–109.
- ZHOU, C., YIN, K., CAO, Y. and AHMED, B. (2016). Application of time series analysis and PSO–SVM model in predicting the Bazimen landslide in the Three Gorges Reservoir, China. *Eng. Geol.* **204** 108–120.

A RANDOM EFFECTS STOCHASTIC BLOCK MODEL FOR JOINT COMMUNITY DETECTION IN MULTIPLE NETWORKS WITH APPLICATIONS TO NEUROIMAGING

BY SUBHADEEP PAUL¹ AND YUGUO CHEN²

¹*Department of Statistics, The Ohio State University, paul.963@osu.edu*

²*Department of Statistics, University of Illinois at Urbana-Champaign, yuguo@illinois.edu*

To analyze data from multisubject experiments in neuroimaging studies, we develop a modeling framework for joint community detection in a group of related networks that can be considered as a sample from a population of networks. The proposed random effects stochastic block model facilitates the study of group differences and subject-specific variations in the community structure. The model proposes a putative mean community structure, which is representative of the group or the population under consideration but is not the community structure of any individual component network. Instead, the community memberships of nodes vary in each component network with a transition matrix, thus modeling the variation in community structure across a group of subjects. To estimate the quantities of interest, we propose two methods: a variational EM algorithm and a model-free “two-step” method called Co-OSNTF which is based on nonnegative matrix factorization. We also develop a resampling-based hypothesis test for differences between community structure in two populations both at the whole network level and node level. The methodology is applied to the COBRE dataset, a publicly available fMRI dataset from multisubject experiments involving schizophrenia patients. Our methods reveal an overall putative community structure representative of the group as well as subject-specific variations within each of the two groups, healthy controls and schizophrenia patients. The model has good predictive ability for predicting community structure in subjects from the same population but outside the training sample. Using our network level hypothesis tests, we are able to ascertain statistically significant difference in community structure between the two groups, while our node level tests help determine the nodes that are driving the difference.

REFERENCES

- ALEXANDER-BLOCH, A. F., GOGTAY, N., MEUNIER, D., BIRN, R., CLASEN, L., LALONDE, F., LENROOT, R., GIEDD, J. and BULLMORE, E. T. (2010). Disrupted modularity and local connectivity of brain functional networks in childhood-onset schizophrenia. *Front. Syst. Neurosci.* **4** 147. <https://doi.org/10.3389/fnsys.2010.00147>
- ALEXANDER-BLOCH, A., LAMBIOTTE, R., ROBERTS, B., GIEDD, J., GOGTAY, N. and BULLMORE, E. (2012). The discovery of population differences in network community structure: New methods and applications to brain functional networks in schizophrenia. *NeuroImage* **59** 3889–3900.
- BACCO, C. D., POWER, E. A., LARREMORE, D. B. and MOORE, C. (2017). Community detection, link prediction, and layer interdependence in multilayer networks. *Phys. Rev. E* **95** 042317. <https://doi.org/10.1103/PhysRevE.95.042317>
- BARBILLON, P., DONNET, S., LAZEGA, E. and BAR-HEN, A. (2017). Stochastic block models for multiplex networks: An application to a multilevel network of researchers. *J. Roy. Statist. Soc. Ser. A* **180** 295–314. <https://doi.org/10.1111/rssa.12193>
- BASSETT, D. S., WYMBBS, N. F., PORTER, M. A., MUCHA, P. J., CARLSON, J. M. and GRAFTON, S. T. (2011). Dynamic reconfiguration of human brain networks during learning. *Proc. Natl. Acad. Sci. USA* **108** 7641–7646.

- BASSETT, D. S., NELSON, B. G., MUELLER, B. A., CAMCHONG, J. and LIM, K. O. (2012). Altered resting state complexity in schizophrenia. *NeuroImage* **59** 2196–2207.
- BAZZI, M., PORTER, M. A., WILLIAMS, S., McDONALD, M., FENN, D. J. and HOWISON, S. D. (2016). Community detection in temporal multilayer networks, with an application to correlation networks. *Multiscale Model. Simul.* **14** 1–41. [MR3439753 https://doi.org/10.1137/15M1009615](https://doi.org/10.1137/15M1009615)
- 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](https://doi.org/10.1137/15M1009615)
- BETZEL, R. F., BERTOLERO, M. A., GORDON, E. M., GRATTON, C., DOSENBACH, N. U. F. and BASSETT, D. S. (2019). The community structure of functional brain networks exhibits scale-specific patterns of inter- and intra-subject variability. *NeuroImage* **202** 115990. <https://doi.org/10.1016/j.neuroimage.2019.07.003>
- BICKEL, P., CHOI, D., CHANG, X. and ZHANG, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.* **41** 1922–1943. [MR3127853 https://doi.org/10.1214/13-AOS1124](https://doi.org/10.1214/13-AOS1124)
- BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R. and LEFEBVRE, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008** P10008.
- BOCCALETTI, S., BIANCONI, G., CRIADO, R., DEL GENIO, C. I., GÓMEZ-GARDEÑES, J., ROMANCE, M., SENDIÑA-NADAL, I., WANG, Z. and ZANIN, M. (2014). The structure and dynamics of multilayer networks. *Phys. Rep.* **544** 1–122. [MR3270140 https://doi.org/10.1016/j.physrep.2014.07.001](https://doi.org/10.1016/j.physrep.2014.07.001)
- BRAUN, U., SCHÄFER, A., WALTER, H., ERK, S., ROMANCZUK-SEIFERTH, N., HADDAD, L., SCHWEIGER, J. I., GRIMM, O., HEINZ, A. et al. (2015). Dynamic reconfiguration of frontal brain networks during executive cognition in humans. *Proc. Natl. Acad. Sci. USA* **112** 11678–11683.
- BROWNLEES, C. T., GUDMUNDSSON, G. and LUGOSI, G. (2017). Community detection in partial correlation network models. Available at SSRN 2776505.
- BULLMORE, E. and SPORNS, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev., Neurosci.* **10** 186–198.
- CELISSE, A., DAUDIN, J.-J. and PIERRE, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.* **6** 1847–1899. [MR2988467 https://doi.org/10.1214/12-EJS729](https://doi.org/10.1214/12-EJS729)
- CHEN, S., KANG, J., XING, Y. and WANG, G. (2015). A parsimonious statistical method to detect groupwise differentially expressed functional connectivity networks. *Hum. Brain Mapp.* **36** 5196–5206.
- 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](https://doi.org/10.1007/s11222-007-9046-7)
- DING, C., LI, T., PENG, W. and PARK, H. (2006). Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 126–135. ACM, New York.
- DONG, X., FROSSARD, P., VANDERGHEYNST, P. and NEFEDOV, N. (2014). Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds. *IEEE Trans. Signal Process.* **62** 905–918. [MR3160322 https://doi.org/10.1109/TSP.2013.2295553](https://doi.org/10.1109/TSP.2013.2295553)
- EDELMAN, A., ARIAS, T. A. and SMITH, S. T. (1999). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20** 303–353. [MR1646856 https://doi.org/10.1137/S0895479895290954](https://doi.org/10.1137/S0895479895290954)
- FORNITO, A., ZALESKY, A. and BREAKSPEAR, M. (2013). Graph analysis of the human connectome: Promise, progress, and pitfalls. *NeuroImage* **80** 426–444.
- FORNITO, A., ZALESKY, A. and BREAKSPEAR, M. (2015). The connectomics of brain disorders. *Nat. Rev., Neurosci.* **16** 159–172.
- FUJITA, A., TAKAHASHI, D. Y., PATRIOTA, A. G. and SATO, J. R. (2014). A non-parametric statistical test to compare clusters with applications in functional magnetic resonance imaging data. *Stat. Med.* **33** 4949–4962. [MR3276511 https://doi.org/10.1002/sim.6292](https://doi.org/10.1002/sim.6292)
- GADDELKARIM, J. J., SCHONFELD, D., AJILORE, O., ZHAN, L., ZHANG, A. F., FEUSNER, J. D., THOMPSON, P. M., SIMON, T. J., KUMAR, A. et al. (2012). A framework for quantifying node-level community structure group differences in brain connectivity networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 196–203. Springer, Berlin.
- GINESTET, C. E., FOURNEL, A. P. and SIMMONS, A. (2014). Statistical network analysis for functional MRI: Summary networks and group comparisons. *Front. Comput. Neurosci.* **8** 51. <https://doi.org/10.3389/fncom.2014.00051>
- GINESTET, C. E., LI, J., BALACHANDRAN, P., ROSENBERG, S. and KOLACZYK, E. D. (2017). Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.* **11** 725–750. [MR3693544 https://doi.org/10.1214/16-AOAS1015](https://doi.org/10.1214/16-AOAS1015)
- GLEREAN, E., PAN, R. K., SALMI, J., KUJALA, R., LAHNAKOSKI, J. M., ROINE, U., NUMMENMAA, L., LEPPÄMÄKI, S., NIEMINEN-VON WENDT, T. et al. (2016). Reorganization of functionally connected brain subnetworks in high-functioning autism. *Hum. Brain Mapp.* **37** 1066–1079.

- HAN, Q., XU, K. S. and AIROLDI, E. M. (2015). Consistent estimation of dynamic and multi-layer block models. In *Proceedings of the 32nd International Conference on Machine Learning* 1511–1520.
- HE, Y., WANG, J., WANG, L., CHEN, Z. J., YAN, C., YANG, H., TANG, H., ZHU, C., GONG, Q. et al. (2009). Uncovering intrinsic modular organization of spontaneous brain activity in humans. *PLoS ONE* **4** e5226.
- HUTCHISON, R. M., WOMELSDORF, T., ALLEN, E. A., BANDETTINI, P. A., CALHOUN, V. D., CORBETTA, M., DELLA PENNA, S., DUYN, J. H., GLOVER, G. H. et al. (2013). Dynamic functional connectivity: Promise, issues, and interpretations. *NeuroImage* **80** 360–378.
- JONES, D. T., VEMURI, P., MURPHY, M. C., GUNTER, J. L., SENJEM, M. L., MACHULDA, M. M., PRZYBELSKI, S. A., GREGG, B. E., KANTARCI, K. et al. (2012). Non-stationarity in the “resting brain’s” modular architecture. *PLoS ONE* **7** e39731. <https://doi.org/10.1371/journal.pone.0039731>
- KIVELÄ, M., ARENAS, A., BARTHELEMY, M., GLEESON, J. P., MORENO, Y. and PORTER, M. A. (2014). Multilayer networks. *J. Complex Netw.* **2** 203–271.
- KUHN, H. W. (1955). The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2** 83–97. [MR0075510 https://doi.org/10.1002/nav.3800020109](https://doi.org/10.1002/nav.3800020109)
- KUJALA, R., GLERAN, E., PAN, R. K., JÄÄSKELÄINEN, I. P., SAMS, M. and SARAMÄKI, J. (2016). Graph coarse-graining reveals differences in the module-level structure of functional brain networks. *Eur. J. Neurosci.* **44** 2673–2684. <https://doi.org/10.1111/ejn.13392>
- KUMAR, A., RAI, P. and DAUME, H. (2011). Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems* 1413–1421.
- LEE, D. D. and SEUNG, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems* 556–562.
- 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](https://doi.org/10.1214/14-AOS1274)
- LIU, Y., LIANG, M., ZHOU, Y., HE, Y., HAO, Y., SONG, M., YU, C., LIU, H., LIU, Z. et al. (2008). Disrupted small-world networks in schizophrenia. *Brain* **131** 945–961.
- LIU, J., WANG, C., GAO, J. and HAN, J. (2013). Multi-view clustering via joint nonnegative matrix factorization. In *Proc. of SDM* **13** 252–260. SIAM, Philadelphia.
- LYNALL, M.-E., BASSETT, D. S., KERWIN, R., MCKENNA, P. J., KITZBICHLER, M., MULLER, U. and BULLMORE, E. (2010). Functional connectivity and brain networks in schizophrenia. *J. Neurosci.* **30** 9477–9487.
- MANKAD, S. and MICHAELIDIS, G. (2013). Structural and functional discovery in dynamic networks with non-negative matrix factorization. *Phys. Rev. E* **88** 042812.
- 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](https://doi.org/10.1111/rssb.12200)
- 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>
- MIRZAL, A. (2014). A convergent algorithm for orthogonal nonnegative matrix factorization. *J. Comput. Appl. Math.* **260** 149–166. [MR3133336 https://doi.org/10.1016/j.cam.2013.09.022](https://doi.org/10.1016/j.cam.2013.09.022)
- MOUSSA, M. N., STEEN, M. R., LAURIENTI, P. J. and HAYASAKA, S. (2012). Consistency of network modules in resting-state FMRI connectome data. *PLoS ONE* **7** e44428. <https://doi.org/10.1371/journal.pone.0044428>
- MUCHA, P. J., RICHARDSON, T., MACON, K., PORTER, M. A. and ONNELA, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science* **328** 876–878. [MR2662590 https://doi.org/10.1126/science.1184819](https://doi.org/10.1126/science.1184819)
- NARAYAN, M. and ALLEN, G. I. (2016). Mixed effects models for resampled network statistics improves statistical power to find differences in multi-subject functional connectivity. *Front. Neurosci.* **10** 108. <https://doi.org/10.3389/fnins.2016.00108>
- NICOSIA, V. and LATORA, V. (2015). Measuring and modeling correlations in multiplex networks. *Phys. Rev. E* **92** 032805.
- PAPADIMITRIOU, C. H. (2003). *Computational Complexity*. Wiley, New York.
- PAUL, S. and CHEN, Y. (2016a). Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electron. J. Stat.* **10** 3807–3870. [MR3579677 https://doi.org/10.1214/16-EJS1211](https://doi.org/10.1214/16-EJS1211)
- PAUL, S. and CHEN, Y. (2016b). Null models and modularity based community detection in multi-layer networks. Preprint. Available at [arXiv:1608.00623](https://arxiv.org/abs/1608.00623).
- PAUL, S. and CHEN, Y. (2016c). Orthogonal symmetric non-negative matrix factorization under the stochastic block model. Preprint. Available at [arXiv:1605.05349](https://arxiv.org/abs/1605.05349).
- PAUL, S. and CHEN, Y. (2020a). Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *Ann. Statist.* **48** 230–250. [MR4065160 https://doi.org/10.1214/18-AOS1800](https://doi.org/10.1214/18-AOS1800)
- PAUL, S. and CHEN, Y. (2020b). Supplement to “A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging.” <https://doi.org/10.1214/20-AOAS1339SUPP>

- PEIXOTO, T. P. (2015). Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Phys. Rev. E* **92** 042807.
- PENNY, W. D., FRISTON, K. J., ASHBURNER, J. T., KIEBEL, S. J. and NICHOLS, T. E. (2011). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press, San Diego.
- PERCIVAL, D. B. and WALDEN, A. T. (2006). *Wavelet Methods for Time Series Analysis. Cambridge Series in Statistical and Probabilistic Mathematics* **4**. Cambridge Univ. Press, Cambridge. [MR2218866](#)
- POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M. et al. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.
- QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems* 3120–3128.
- REICHARDT, J. and BORNHOLDT, S. (2006). Statistical mechanics of community detection. *Phys. Rev. E* (3) **74** 016110, 14. [MR2276596](#) <https://doi.org/10.1103/PhysRevE.74.016110>
- REYES, P. and RODRIGUEZ, A. (2016). Stochastic blockmodels for exchangeable collections of networks. Preprint. Available at [arXiv:1606.05277](#).
- ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic block-model. *Ann. Statist.* **39** 1878–1915. [MR2893856](#) <https://doi.org/10.1214/11-AOS887>
- RUBINOV, M. and SPORNS, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage* **52** 1059–1069.
- SIMPSON, S. L., BOWMAN, F. D. and LAURIENTI, P. J. (2013). Analyzing complex functional brain networks: Fusing statistics and network science to understand the brain. *Stat. Surv.* **7** 1–36. [MR3161730](#) <https://doi.org/10.1214/13-SS103>
- SIMPSON, S. L., LYDAY, R. G., HAYASAKA, S., MARSH, A. P. and LAURIENTI, P. J. (2013). A permutation testing framework to compare groups of brain networks. *Front. Comput. Neurosci.* **7** 171. <https://doi.org/10.3389/fncom.2013.00171>
- SPORNS, O. (2014). Contributions and challenges for network models in cognitive neuroscience. *Nat. Neurosci.* **17** 652–660.
- STAM, C. J. (2014). Modern network science of neurological disorders. *Nat. Rev., Neurosci.* **15** 683–695.
- STANLEY, N., SHAI, S., TAYLOR, D. and MUCHA, P. J. (2016). Clustering network layers with the strata multilayer stochastic block model. *IEEE Trans. Netw. Sci. Eng.* **3** 95–105. [MR3515211](#) <https://doi.org/10.1109/TNSE.2016.2537545>
- STEEN, M., HAYASAKA, S., JOYCE, K. and LAURIENTI, P. (2011). Assessing the consistency of community structure in complex networks. *Phys. Rev. E* **84** 016111.
- STEVENS, A. A., TAPPON, S. C., GARG, A. and FAIR, D. A. (2012). Functional brain network modularity captures inter- and intra-individual variation in working memory capacity. *PLoS ONE* **7** e30468. <https://doi.org/10.1371/journal.pone.0030468>
- STEWART, G. W. and SUN, J. G. (1990). *Matrix Perturbation Theory. Computer Science and Scientific Computing*. Academic Press, Boston, MA. [MR1061154](#)
- SWEET, T. M., THOMAS, A. C. and JUNKER, B. W. (2015). Hierarchical mixed membership stochastic blockmodels for multiple networks and experimental interventions. In *Handbook of Mixed Membership Models and Their Applications. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 463–487. CRC Press, Boca Raton, FL. [MR3380043](#)
- TANG, W., LU, Z. and DHILLON, I. S. (2009). Clustering with multiple graphs. In *Proceedings of the Ninth IEEE International Conference on Data Mining* 1016–1021. IEEE Press, New York.
- TZOURIO-MAZOYER, N., LANDEAU, B., PAPATHANASSIOU, D., CRIVELLO, F., ETARD, O., DELCROIX, N., MAZOYER, B. and JOLIOT, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15** 273–289.
- VALLES-CATALA, T., MASSUCCI, F. A., GUIMERA, R. and SALES-PARDO, M. (2016). Multilayer stochastic block models reveal the multilayer structure of complex networks. *Phys. Rev. X* **6** 011036.
- VAN DEN HEUVEL, M. P. and POL, H. E. H. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.* **20** 519–534.
- VAN DEN HEUVEL, M. P., MANDL, R. C., STAM, C. J., KAHN, R. S. and POL, H. E. H. (2010). Aberrant frontal and temporal complex network structure in schizophrenia: A graph theoretical analysis. *J. Neurosci.* **30** 15915–15926.
- VAN DEN HEUVEL, M. P., SPORNS, O., COLLIN, G., SCHEEWE, T., MANDL, R. C., CAHN, W., GOÑI, J., POL, H. E. H. and KAHN, R. S. (2013). Abnormal rich club organization and functional brain dynamics in schizophrenia. *JAMA Psychiatry* **70** 783–792.
- WANG, J., ZUO, X., DAI, Z., XIA, M., ZHAO, Z., ZHAO, X., JIA, J., HAN, Y. and HE, Y. (2013). Disrupted functional brain connectome in individuals at risk for Alzheimer’s disease. *Biological Psychiatry* **73** 472–481.

- WEBER, M. J., DETRE, J. A., THOMPSON-SCHILL, S. L. and AVANTS, B. B. (2013). Reproducibility of functional network metrics and network structure: A comparison of task-related BOLD, resting ASL with BOLD contrast, and resting cerebral blood flow. *Cognitive, Affective, & Behavioral Neuroscience* **13** 627–640.
- WHITFIELD-GABRIELI, S. and NIETO-CASTANON, A. (2012). CONN: A functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect.* **2** 125–141.
- XIA, M., WANG, J. and HE, Y. (2013). BrainNet viewer: A network visualization tool for human brain connectomics. *PLoS ONE* **8** e68910.
- YU, Q., PLIS, S. M., ERHARDT, E. B., ALLEN, E. A., SUI, J., KIEHL, K. A., PEARLSON, G. and CALHOUN, V. D. (2011). Modular organization of functional network connectivity in healthy controls and patients with schizophrenia during the resting state. *Front. Syst. Neurosci.* **5** 103. <https://doi.org/10.3389/fnsys.2011.00103>
- YU, Q., ALLEN, E. A., SUI, J., ARBABSHIRANI, M. R., PEARLSON, G. and CALHOUN, V. D. (2012). Brain connectivity networks in schizophrenia underlying resting state functional magnetic resonance imaging. *Current Topics in Medicinal Chemistry* **12** 2415–2425.
- ZALESKY, A., FORNITO, A. and BULLMORE, E. T. (2010). Network-based statistic: Identifying differences in brain networks. *NeuroImage* **53** 1197–1207.

ANALYSES OF PREVENTIVE CARE MEASURES WITH INCOMPLETE HISTORICAL DATA IN ELECTRONIC MEDICAL RECORDS: AN EXAMPLE FROM COLORECTAL CANCER SCREENING

BY YINGYE ZHENG¹, DOUGLAS A. CORLEY², CHYKE DOUBENI³, ETHAN HALM⁴,
SUSAN M. SHORTREED⁵, WILLIAM E. BARLOW⁶, ANN ZAUBER⁷, TOR
DEVIN TOSTESON⁸, AND JESSICA CHUBAK⁵

¹Department of Biostatistics, Fred Hutchinson Cancer Research Center, yzheng@fredhutch.org

²Division of Research, Kaiser Permanente Northern California, douglas.corley@kp.org

³Department of Family Medicine, Mayo Clinic School of Medicine, Doubeni.Chyke@mayo.edu

⁴Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical School,
Ethan.Halm@UTSouthwestern.edu

⁵Kaiser Permanente Washington Health Research Institute, Susan.M.Shortreed@kp.org

⁶Cancer Research and Biostatistics, williamb@crab.org

⁷Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, zauber@MSKCC.ORG

⁸Biomedical Data Science, Dartmouth College, tor.tosteson@dartmouth.edu

The calculation of quality of care measures based on electronic medical records (EMRs) may be inaccurate because of incomplete capture of past services. We evaluate the influence of different statistical approaches for calculating the proportion of patients who are up-to-date for a preventive service, using the example of colorectal cancer (CRC) screening. We propose an extension of traditional mixture models to account for the uncertainty in compliance which is further complicated by the choice of various screening modalities with different recommended screening intervals. We conducted simulation studies to compare various statistical approaches and demonstrated that the proposed method can alleviate bias when individuals with complete prior medical history information were not representative of the targeted population. The method is motivated by and applied to data from the National Cancer Institute–funded consortium Population-Based Research Optimizing Screening through Personalized Regiments (PROSPR). Findings from the application are important for the evaluation of appropriate use of preventive care and provide a novel tool for dealing with similar analytical challenges with EMR data in broad settings.

REFERENCES

- BRUNELLI, S. M., GAGNE, J. J., HUYBRECHTS, K. F., WANG, S. V., PATRICK, A. R., ROTHMAN, K. J. and SEEGER, J. D. (2013). Estimation using all available covariate information versus a fixed look-back window for dichotomous covariates. *Pharmacoepidemiol. Drug Saf.* **22** 542–550.
- BURNETT-HARTMAN, A. N., MEHTA, S. J., ZHENG, Y., GHAI, N. R., MCLERRAN, D. F., CHUBAK, J., QUINN, V. P., SKINNER, C. S., CORLEY, D. A. et al. (2016). Racial/ethnic disparities in colorectal cancer screening across healthcare systems. *Am. J. Prev. Med.* **51** e107–e115.
- CHENG, S. C., WEI, L. J. and YING, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82** 835–845. [MR1380818 https://doi.org/10.1093/biomet/82.4.835](https://doi.org/10.1093/biomet/82.4.835)
- CHUBAK, J. and HUBBARD, R. (2016). Defining and measuring adherence to cancer screening. *Journal of Medical Screening* **23** 1–1.
- COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276. [MR0400509 https://doi.org/10.1093/biomet/62.2.269](https://doi.org/10.1093/biomet/62.2.269)
- FAREWELL, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1041–1046.
- HUBBARD, R. A., CHUBAK, J. and RUTTER, C. M. (2014). Estimating screening test utilization using electronic health records data. *EGEMS* **2**.

- HUBBARD, R. A., JOHNSON, E., CHUBAK, J., WERNLI, K. J., KAMINENI, A., BOGART, A. and RUTTER, C. M. (2017). Accounting for misclassification in electronic health records-derived exposures using generalized linear finite mixture models. *Health Serv. Outcomes Res. Methodol.* **17** 101–112. <https://doi.org/10.1007/s10742-016-0149-5>
- KLABUNDE, C. N., ZHENG, Y., QUINN, V. P., BEABER, E. F., RUTTER, C. M., HALM, E. A., CHUBAK, J., DOUBENI, C. A., HAAS, J. S. et al. (2016). Influence of age and comorbidity on colorectal cancer screening in the elderly. *Am. J. Prev. Med.* **51** e67–e75.
- KLEINBAUM, D. G. and KLEIN, M. (2002). Maximum likelihood techniques: An overview. *Logistic Regression: A Self-Learning Text* 101–124.
- KOHN, L. T., CORRIGAN, J. M., DONALDSON, M. S. et al. (2000). To err is human: Building a safer health system. a report of the committee on quality of health care in America, Institute of Medicine.
- LANDON, B. E., O'MALLEY, A. J. and KEEGAN, T. (2010). Can choice of the sample population affect perceived performance: Implications for performance assessment. *J. Gen. Intern. Med.* **25** 104–109.
- LIN, J. S., PIPER, M. A., PERDUE, L. A., RUTTER, C. M., WEBBER, E. M., O'CONNOR, E., SMITH, N. and WHITLOCK, E. P. (2016). Screening for colorectal cancer: Updated evidence report and systematic review for the us preventive services task force. *JAMA* **315** 2576–2594.
- LITTLE, R. J. A. and RUBIN, D. B. (2014). *Statistical Analysis with Missing Data*. Wiley, Hoboken, NJ.
- RUTTER, C. M., GREENLEE, R. T., JOHNSON, E., STARK, A., WEINMANN, S., KAMINENI, A., ADAMS, K. and DOUBENI, C. A. (2015). Prevalence of colonoscopy before age 50. *Prev. Med.* **72** 126–129.
- SHORTREED, S., JOHNSON, E., RUTTER, C., KAMINEI, A., WERNLI, K. and CHUBAK, J. (2016). Cohort restriction based on prior enrollment: Examining potential biases in estimating cancer and mortality risk. *Observational Studies.* **2** 51–64.
- SY, J. P. and TAYLOR, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics* **56** 227–236. [MR1767631 https://doi.org/10.1111/j.0006-341X.2000.00227.x](https://doi.org/10.1111/j.0006-341X.2000.00227.x)
- TIRO, J. A., KAMINENI, A., LEVIN, T. R., ZHENG, Y., SCHOTTINGER, J. S., RUTTER, C. M., CORLEY, D. A., SKINNER, C. S., CHUBAK, J. et al. (2014). The colorectal cancer screening process in community settings: A conceptual model for the population-based research optimizing screening through personalized regimens consortium. *Cancer Epidemiol. Biomark. Prev.* **23** 1147–1158.
- ZHENG, Y., CORLEY, D. A., DOUBENI, C., HALM, E., SHORTREED, S. M., BARLOW, W. E., ZAUBER, A., TOSTESON, T. D., and CHUBAK, J., (2020). Supplement to “Analyses of preventive care measures with incomplete historical data in electronic medical records: An example from colorectal cancer screening.” <https://doi.org/10.1214/20-AOAS1342SUPP>

ACCOUNTING FOR DEPENDENT ERRORS IN PREDICTORS AND TIME-TO-EVENT OUTCOMES USING ELECTRONIC HEALTH RECORDS, VALIDATION SAMPLES AND MULTIPLE IMPUTATION

BY MARK J. GIGANTI^{1,*}, PAMELA A. SHAW², GUANHUA CHEN³,
SALLY S. BEBAWY^{4,‡}, MEGAN M. TURNER^{4,§}, TIMOTHY R. STERLING^{4,¶} AND
BRYAN E. SHEPHERD^{1,†}

¹Department of Biostatistics, Vanderbilt University, *mark.giganti@vanderbilt.edu; †bryan.shepherd@vanderbilt.edu

²Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, shawp@penndmedicine.upenn.edu

³Department of Biostatistics and Medical Informatics, University of Wisconsin, gchen25@wisc.edu

⁴Vanderbilt University School of Medicine, ‡sally.furukawa@vumc.org; §megan.turner@vumc.org;
¶timothy.sterling@vanderbilt.edu

Data from electronic health records (EHR) are prone to errors which are often correlated across multiple variables. The error structure is further complicated when analysis variables are derived as functions of two or more error-prone variables. Such errors can substantially impact estimates, yet we are unaware of methods that simultaneously account for errors in covariates and time-to-event outcomes. Using EHR data from 4217 patients, the hazard ratio for an AIDS-defining event associated with a 100 cell/mm³ increase in CD4 count at ART initiation was 0.74 (95%CI: 0.68–0.80) using unvalidated data and 0.60 (95%CI: 0.53–0.68) using fully validated data. Our goal is to obtain unbiased and efficient estimates after validating a random subset of records. We propose fitting discrete failure time models to the validated subsample and then multiply imputing values for unvalidated records. We demonstrate how this approach simultaneously addresses dependent errors in predictors, time-to-event outcomes, and inclusion criteria. Using the fully validated dataset as a gold standard, we compare the mean squared error of our estimates with those from the unvalidated dataset and the corresponding subsample-only dataset for various subsample sizes. By incorporating reasonably sized validated subsamples and appropriate imputation models, our approach had improved estimation over both the naive analysis and the analysis using only the validation subsample.

REFERENCES

- BURTON, A., ALTMAN, D. G., ROYSTON, P. and HOLDER, R. L. (2006). The design of simulation studies in medical statistics. *Stat. Med.* **25** 4279–4292. [MR2307592 https://doi.org/10.1002/sim.2673](https://doi.org/10.1002/sim.2673)
- CARPENTER, J. R., KENWARD, M. G. and VANSTEELENDT, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J. Roy. Statist. Soc. Ser. A* **169** 571–584. [MR2236921 https://doi.org/10.1111/j.1467-985X.2006.00407.x](https://doi.org/10.1111/j.1467-985X.2006.00407.x)
- CHAN, K. S., FOWLES, J. B. and WEINER, J. P. (2010). Electronic health records and reliability and validity of quality measures: A review of the literature. *Med. Care Res. Rev.* **67** 742–752.
- COLE, S. R., CHU, H. and GREENLAND, S. (2006). Multiple-imputation for measurement-error correction. *Int. J. Epidemiol.* **35** 1074–1081.
- COOK, J. R. and STEFANSKI, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* **89** 1314–1328.
- D’AGOSTINO, R. B., LEE, M.-L., BELANGER, A. J., CUPPLES, L. A., ANDERSON, K. and KANNEL, W. B. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham Heart Study. *Stat. Med.* **9** 1501–1515.
- DUDA, S. N., SHEPHERD, B. E., GADD, C. S., MASYS, D. R. and MCGOWAN, C. C. (2012). Measuring the quality of observational study data in an international HIV research network. *PLoS ONE* **7** e33908.

Key words and phrases. Correlated errors, electronic health records, measurement error, multiple imputation, survival analysis.

- EDWARDS, J. K., COLE, S. R., TROESTER, M. A. and RICHARDSON, D. B. (2013). Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *Am. J. Epidemiol.* **177** 904–912.
- EFRON, B. (1988). Logistic regression, survival analysis, and the Kaplan–Meier curve. *J. Amer. Statist. Assoc.* **83** 414–425. [MR0971367](#)
- FLOYD, J. S., HECKBERT, S. R., WEISS, N. S., CARRELL, D. S. and PSATY, B. M. (2012). Use of administrative data to estimate the incidence of statin-related rhabdomyolysis. *J. Am. Med. Assoc.* **307** 1580–1582.
- GIGANTI, M. J., SHAW, P. A., CHEN, G., BEBAWY, S. S., TURNER, M. M., STERLING, T. R. and SHEPHERD, B. E. (2020). Supplement to “Accounting for dependent errors in predictors and time-to-event outcomes using electronic health records, validation samples and multiple imputation.” <https://doi.org/10.1214/20-AOAS1343SUPP>
- HERNÁN, M. A., BRUMBACK, B. and ROBINS, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J. Amer. Statist. Assoc.* **96** 440–448. [MR1939347](#) <https://doi.org/10.1198/016214501753168154>
- HUANG, Y. and WANG, C. Y. (2000). Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. *J. Amer. Statist. Assoc.* **95** 1209–1219. [MR1804244](#) <https://doi.org/10.2307/2669761>
- HUNSBERGER, S., ALBERT, P. S. and DODD, L. (2010). Analysis of progression-free survival data using a discrete time survival model that incorporates measurements with and without diagnostic error. *Clin. Trials* **7** 634–642.
- KOEHLER, E., BROWN, E. and HANEUSE, S. J.-P. A. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *Amer. Statist.* **63** 155–162. [MR2750076](#) <https://doi.org/10.1198/tast.2009.0030>
- KORN, E. L., DODD, L. E. and FREIDLIN, B. (2010). Measurement error in the timing of events: Effect on survival analyses in randomized clinical trials. *Clin. Trials* **7** 626–633.
- LI, Y. and LIN, X. (2003). Functional inference in frailty measurement error models for clustered survival data using the SIMEX approach. *J. Amer. Statist. Assoc.* **98** 191–203. [MR1965685](#) <https://doi.org/10.1198/016214503388619210>
- LITTLE, R. J. and RUBIN, D. B. (2014). *Statistical Analysis with Missing Data* **333**. John Wiley & Sons.
- LUMLEY, T., SHAW, P. A. and DAI, J. Y. (2011). Connections between survey calibration estimators and semi-parametric models for incomplete data. *Int. Stat. Rev.* **79** 200–220. <https://doi.org/10.1111/j.1751-5823.2011.00138.x>
- MAGARET, A. S. (2008). Incorporating validation subsets into discrete proportional hazards models for mismeasured outcomes. *Stat. Med.* **27** 5456–5470. [MR2542363](#) <https://doi.org/10.1002/sim.3365>
- MCCLINTOCK, B. T., JOHNSON, D. S., HOOTEN, M. B., HOEF, J. M. V. and MORALES, J. M. (2014). When to be discrete: The importance of time formulation in understanding animal movement. *Mov. Ecol.* **2** 21. <https://doi.org/10.1186/s40462-014-0021-6>
- MCISAAC, M. and COOK, R. J. (2017). Statistical methods for incomplete data: Some results on model misspecification. *Stat. Methods Med. Res.* **26** 248–267. [MR3592724](#) <https://doi.org/10.1177/0962280214544251>
- NAKAMURA, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika* **77** 127–137. [MR1049414](#) <https://doi.org/10.1093/biomet/77.1.127>
- OH, E. J., SHEPHERD, B. E., LUMLEY, T. and SHAW, P. A. (2018). Considerations for analysis of time-to-event outcomes measured with error: Bias and correction with SIMEX. *Stat. Med.* **37** 1276–1289. [MR3777974](#) <https://doi.org/10.1002/sim.7554>
- PRENTICE, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69** 331–342. [MR0671971](#) <https://doi.org/10.1093/biomet/69.2.331>
- RICHARDSON, B. A. and HUGHES, J. P. (2000). Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. *Biostatistics* **1** 341–354.
- ROBINS, J. M. and WANG, N. (2000). Inference for imputation estimators. *Biometrika* **87** 113–124. [MR1766832](#) <https://doi.org/10.1093/biomet/87.1.113>
- SHAW, P. A. and PRENTICE, R. L. (2012). Hazard ratio estimation for biomarker-calibrated dietary exposures. *Biometrics* **68** 397–407. [MR2959606](#) <https://doi.org/10.1111/j.1541-0420.2011.01690.x>
- SHEPHERD, B. E., SHAW, P. A. and DODD, L. E. (2012). Using audit information to adjust parameter estimates for data errors in clinical trials. *Clin. Trials* **9** 721–729.
- SHEPHERD, B. E. and YU, C. (2011). Accounting for data errors discovered from an audit in multiple linear regression. *Biometrics* **67** 1083–1091. [MR2829243](#) <https://doi.org/10.1111/j.1541-0420.2010.01543.x>
- SKINNER, C. J. and HUMPHREYS, K. (1999). Weibull regression for lifetimes measured with error. *Lifetime Data Anal.* **5** 23–37. [MR1750340](#) <https://doi.org/10.1023/A:1009674915476>
- TSIATIS, A. A. and DAVIDIAN, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88** 447–458. [MR1844844](#) <https://doi.org/10.1093/biomet/88.2.447>

- TURCHIN, P. (1998). *Quantitative Analysis of Movement: Measuring and Modeling Population Redistribution in Animals and Plants* **1**. Sinauer Associates Sunderland.
- TUTZ, G. and SCHMID, M. (2016). *Modeling Discrete Time-to-Event Data*. *Springer Series in Statistics*. Springer, Cham. [MR3497009](#) <https://doi.org/10.1007/978-3-319-28158-2>
- VAN BUUREN, S. (2018). *Flexible Imputation of Missing Data*. CRC Press/CRC.
- 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>

The Annals of Applied Statistics

Next Issues

- Causal inference from observational studies with clustered interference, with application to a cholera vaccine study BRIAN G. BARKLEY, MICHAEL G. HUDGENS, JOHN D. CLEMENS, MOHAMMAD ALI AND MICHAEL E. EMCH
- Quantifying time-varying sources in magnetoencephalography—a discrete approach
ZHIGANG YAO, ZENGYAN FAN, MASAHITO HAYASHI AND WILLIAM EDDY
- Active matrix factorization for surveys
CHELSEA ZHANG, SEAN J. TAYLOR, CURTISS COBB AND JASJEET SEKHON
- Optimal EMG placement for a robotic prosthesis controller with sequential, adaptive functional estimation JONATHAN STALLRICH, MD NAZMUL ISLAM, ANA-MARIA STAICU, DUSTIN CROUCH, LIZHI PAN AND HE HUANG
- Size estimation of key populations in the HIV epidemic in Eswatini using incomplete and misaligned capture-recapture data ABHIRUP DATTA ANDREW PITA, AMRITA RAO, BHEKIE SITHOLE, ZANDILE MNISI AND STEFAN BARAL
- A Bayesian hierarchical model for evaluating forensic footwear evidence
NEIL A. SPENCER AND JARED S. MURRAY
- Statistical methods for replicability assessment KENNETH HUNG AND WILLIAM FITHIAN
- Statistical methods for analysis of combined categorical biomarker data from multiple studies
CHAO CHENG AND MOLIN WANG
- Does terrorism trigger online hate speech? On the association of events and time series
ERIK SCHARWÄCHTER AND EMMANUEL MÜLLER
- A semiparametric mixture method for local false discovery rate estimation from multiple studies
SEOK-OH JENG, DONGSEOK CHOI AND WONCHEOL JANG
- Mining events with declassified diplomatic documents
YUANJUN GAO, JACK GOETZ, MATTHEW CONNELLY AND RAHUL MAZUMDER
- A novel change point approach for the detection of gas emission sources using remotely contained concentration data IDRIS ECKLEY, CLAUDIA KIRCH AND SILKE WEBER
- PTEM: A popularity-based topical expertise model for community question answering
HOHYUN JUNG, JAE-GIL LEE, NAMGIL LEE AND SUNG-HO KIM
- Spatio-temporal probabilistic wind vector forecasting over Saudi Arabia
AMANDA LENZI AND MARC G. GENTON
- Markov decision processes with dynamic transition probabilities: An analysis of shooting strategies in basketball NATHAN SANDHOLTZ AND LUKE BORN
- The Jensen effect and functional single index models: Estimating the ecological implications of nonlinear reaction norms ZI YE, GILES HOOKER AND STEPHEN P. ELLNER
- Efficiency in lung transplant allocation strategies
JINGJING ZOU, DAVID J. LEDERER, DANIEL RABINOWITZ
- Monotonic effects of characteristics on returns
JARED DALE FISHER, DAVID WALKER PUELZ AND CARLOS MARINHO CARVALHO

Continued

The Annals of Applied Statistics

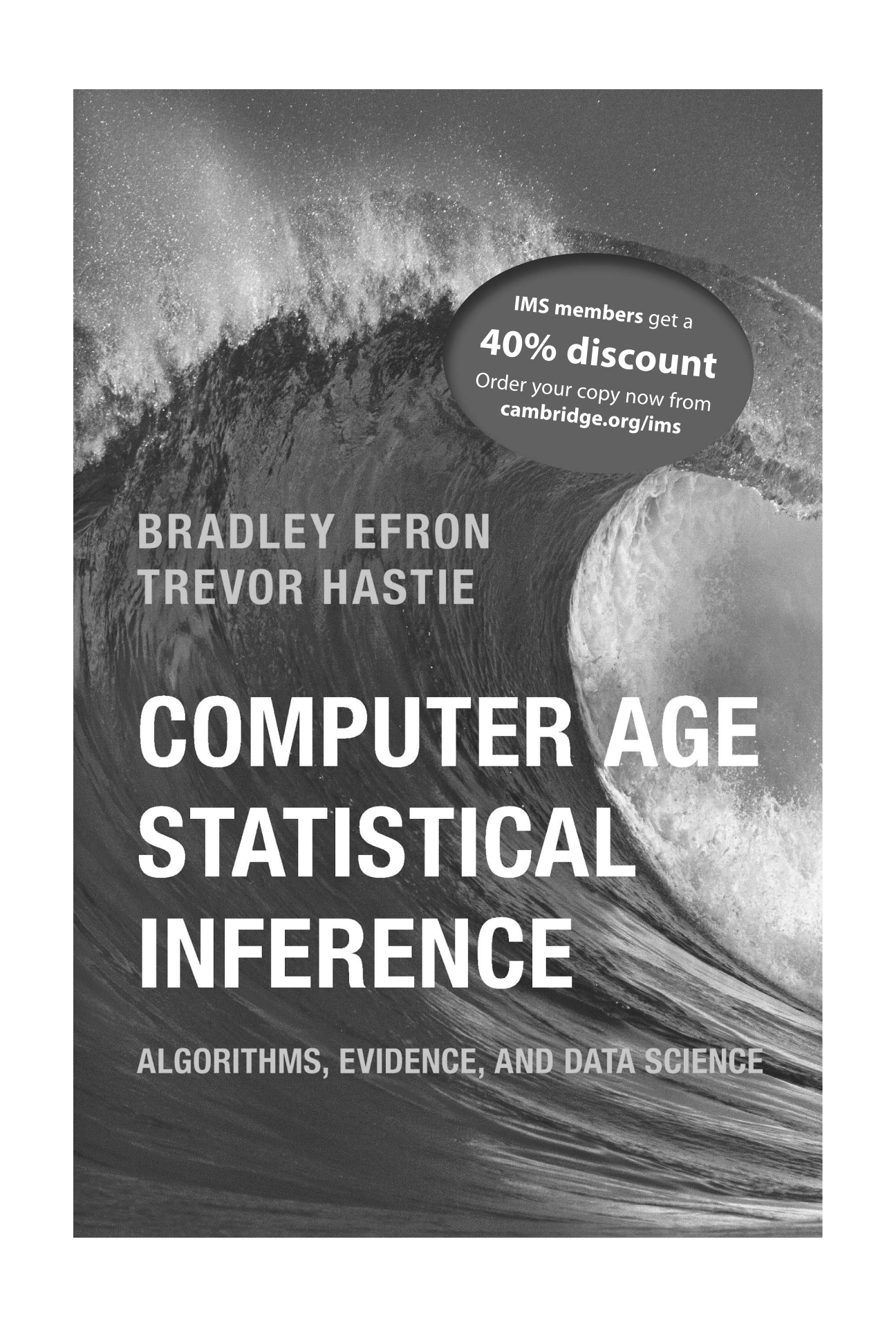
Next Issues—Continued

Hawkes binomial topic model with applications to coupled conflict-Twitter data

GEORGE MOHLER, ERIN MCGRATH, CODY BUNTAIN AND GARY LAFREE

Effective model calibration via sensible variable identification and adjustment, with application to composite fuselage simulation YAN WANG, XIAOWEI YUE, RUI TUO, JEFFREY HUNT AND JIANJUN SHI

A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes . . . MATTHEW KOSLOVSKY, KRISTI L. HOFFMAN, CARRIE R. DANIEL AND MARINA VANNUCCI



IMS members get a
40% discount

Order your copy now from
cambridge.org/ims

BRADLEY EFRON
TREVOR HASTIE

COMPUTER AGE STATISTICAL INFERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE