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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles

continued

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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles—*Continued from front cover*

Correction

THE ANNALS OF APPLIED STATISTICS Vol. 14, No. 1, pp. 1–520 March 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, Heidelberg Institute for Theoretical Studies, 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*: Cun-Hui Zhang, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, USA

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 1, March 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.

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 1–27
https://doi.org/10.1214/19-AOAS1265

SHOPPER: A PROBABILISTIC MODEL OF CONSUMER CHOICE WITH SUBSTITUTES AND COMPLEMENTS

BY FRANCISCO J. R. RUIZ[1], SUSAN ATHEY[2] AND DAVID M. BLEI[3]

[1]*Department of Engineering, University of Cambridge, f.ruiz@columbia.edu*
[2]*Stanford Graduate School of Business, Stanford University, athey@susanathey.com*
[3]*Department of Statistics, Department of Computer Science, Columbia Data Science Institute, Columbia University, david.blei@columbia.edu*

We develop SHOPPER, a sequential probabilistic model of shopping data. SHOPPER uses interpretable components to model the forces that drive how a customer chooses products; in particular, we designed SHOPPER to capture how items interact with other items. We develop an efficient posterior inference algorithm to estimate these forces from large-scale data, and we analyze a large dataset from a major chain grocery store. We are interested in answering counterfactual queries about changes in prices. We found that SHOPPER provides accurate predictions even under price interventions, and that it helps identify complementary and substitutable pairs of products.

REFERENCES

ABERNETHY, J., BACH, F., EVGENIOU, T. and VERT, J. P. (2009). A new approach to collaborative filtering: Operator estimation with spectral regularization. *J. Mach. Learn. Res.* **10** 803–826.

ARORA, S., LI, Y., LIANG, Y. and MA, T. (2016). RAND-WALK: A latent variable model approach to word embeddings. *Transact. Assoc. Comput. Linguist.* **4**.

ATHEY, S. and STERN, S. (1998). An empirical framework for testing theories about complimentarity in organizational design. Technical report, National Bureau of Economic Research, Cambridge, MA.

BAMLER, R. and MANDT, S. (2017). Dynamic word embeddings via skip-gram filtering. In *International Conference in Machine Learning*.

BARKAN, O. (2016). Bayesian neural word embedding. Preprint. Available at arXiv:1603.06571.

BARKAN, O. and KOENIGSTEIN, N. (2016). Item2Vec: Neural item embedding for collaborative filtering. In *IEEE International Workshop on Machine Learning for Signal Processing*.

BENGIO, Y., DUCHARME, R., VINCENT, P. and JANVIN, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.* **3** 1137–1155.

BENGIO, Y., SCHWENK, H., SENÉCAL, J. S., MORIN, F. and GAUVAIN, J. L. (2006). Neural probabilistic language models. In *Innovations in Machine Learning* Springer, Berlin.

BERRY, S. (2014). Structural models of complementary choices. *Mark. Lett.* **25** 245–256.

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

BLUM, J. R. (1954). Approximation methods which converge with probability one. *Ann. Math. Stat.* **25** 382–386. MR0062399 https://doi.org/10.1214/aoms/1177728794

BOTTOU, L., CURTIS, F. E. and NOCEDAL, J. (2018). Optimization methods for large-scale machine learning. *SIAM Rev.* **60** 223–311. MR3797719 https://doi.org/10.1137/16M1080173

BROWNING, M. and MEGHIR, C. (1991). The effects of male and female labor supply on commodity demands. *Econometrica* **59** 925–951.

CANNY, J. (2004). GaP: A factor model for discrete data. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

CATTELL, R. B. (1952). *Factor Analysis*: *An Introduction and Manual for the Psychologist and Social Scientist*. Harper, New York.

CHE, H., CHEN, X. and CHEN, Y. (2012). Investigating effects of out-of-stock on consumer stockkeeping unit choice. *J. Mark. Res.* **49** 502–513.

CHINTAGUNTA, P. K. (1994). Heterogeneous logit model implications for brand positioning. *J. Mark. Res.* 304–311.

Key words and phrases. Probabilistic modeling, discrete choice model, causal interventions, variational inference, market basket data.

CHINTAGUNTA, P. K., NAIR and HARIKESH, S. (2011). Structural workshop paper—discrete-choice models of consumer demand in marketing. *Mark. Sci.* **30** 977–996.

DEATON, A. and MUELLBAUER, J. (1980). An almost ideal demand system. *Am. Econ. Rev.* **70** 312–326.

DONNELLY, R., RUIZ, F. J. R., BLEI, D. M. and ATHEY, S. (2019). Counterfactual inference for consumer choice across many product categories. Available at arXiv:1906.02635.

DOSHI-VELEZ, F., MILLER, K. T., VAN GAEL, J. and TEH, Y. W. (2009). Variational inference for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics* 12.

ELROD, T. (1988). Choice map: Inferring a product-market map from panel data. *Mark. Sci.* **7** 21–40.

ELROD, T. and KEANE, M. P. (1995). A factor-analytic probit model for representing the market structure in panel data. *J. Mark. Res.* **32** 1–16.

FIRTH, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis* (*Special Volume of the Philological Society*) 1952–1959.

GENTZKOW, M. (2007). Valuing new goods in a model with complementarity: Online newspapers. *Am. Econ. Rev.* **97** 713–744.

GOPALAN, P., HOFMAN, J. and BLEI, D. M. (2015). Scalable recommendation with hierarchical Poisson factorization. In *Uncertainty in Artificial Intelligence* 326–335. AUAI Press, Arlington, VA.

GOPALAN, P., RUIZ, F. J. R., RANGANATH, R. and BLEI, D. M. (2014). Bayesian nonparametric Poisson factorization for recommendation systems. In *Artificial Intelligence and Statistics*.

GÖRÜR, D., JÄKEL, F. and RASMUSSEN, C. E. (2006). A choice model with infinitely many latent features. In *International Conference on Machine Learning*.

HARRIS, Z. S. (1954). Distributional structure. *Word* **10** 146–162.

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

HOTZ, V. J. and MILLER, R. A. (1993). Conditional choice probabilities and the estimation of dynamic models. *Rev. Econ. Stud.* **60** 497–529. MR1236835 https://doi.org/10.2307/2298122

HU, Y., KOREN, Y. and VOLINSKY, C. (2008). Collaborative filtering for implicit feedback datasets. In *IEEE International Conference on Data Mining*.

JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.

KEANE, M. P. et al. (2013). Panel data discrete choice models of consumer demand. Prepared for *The Oxford Handbooks: Panel Data*.

KINGMA, D. P. and WELLING, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

LEVY, O. and GOLDBERG, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*.

LIANG, D., ALTOSAAR, J., CHARLIN, L. and BLEI, D. M. (2015). Factorization meets the item embedding: Regularizing matrix factorization with item co-occurrence. In *ACM Conference on Recommender System*.

MA, H., LIU, C., KING, I. and LYU, M. R. (2011). Probabilistic factor models for web site recommendation. In *Proceedings of the* 34*th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

MIKOLOV, T., YIH, W. T. and ZWEIG, G. (2013). Linguistic regularities in continuous space word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S. and DEAN, J. (2013a). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*.

MIKOLOV, T., CHEN, K., CORRADO, G. S. and DEAN, J. (2013b). Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

MNIH, A. and HINTON, G. E. (2007). Three new graphical models for statistical language modelling. In *International Conference on Machine Learning*.

MNIH, A. and KAVUKCUOGLU, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems*.

MNIH, A. and TEH, Y. W. (2012). A fast and simple algorithm for training neural probabilistic language models. In *International Conference on Machine Learning*.

NAESSETH, C., RUIZ, F. J. R., LINDERMAN, S. and BLEI, D. M. (2017). Reparameterization gradients through acceptance-rejection methods. In *Artificial Intelligence and Statistics*.

NG, A. Y. and RUSSELL, S. J. (2000). Algorithms for inverse reinforcement learning. In *International Conference in Machine Learning*.

PENNINGTON, J., SOCHER, R. and MANNING, C. D. (2014). GloVe: Global vectors for word representation. In *Conference on Empirical Methods on Natural Language Processing*.

REZENDE, D. J., MOHAMED, S. and WIERSTRA, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*.

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

RUDOLPH, M., RUIZ, F. J. R., MANDT, S. and BLEI, D. M. (2016). Exponential family embeddings. In *Advances in Neural Information Processing Systems*.

RUIZ, F. J., ATHEY, S. and BLEI, D. M. (2020). Supplement to "SHOPPER: A probabilistic model of consumer choice with substitutes and complements." https://doi.org/10.1214/19-AOAS1265SUPP.

RUIZ, F. J. R., TITSIAS, M. K. and BLEI, D. M. (2016). The generalized reparameterization gradient. In *Advances in Neural Information Processing Systems*.

RUIZ, F. J. R., TITSIAS, M. K., DIENG, A. B. and BLEI, D. M. (2018). Augment and reduce: Stochastic inference for large categorical distributions. In *International Conference on Machine Learning*.

RUSSELL, S. J. (1998). Learning agents for uncertain environments. In *Annual Conference on Computational Learning Theory*.

SEMENOVA, V., GOLDMAN, M., CHERNOZHUKOV, V. and TADDY, M. (2018). Orthogonal ML for demand estimation: High dimensional causal inference in dynamic panels. Available at arXiv:1712.09988.

SONG, I. and CHINTAGUNTA, P. K. (2007). A discrete–continuous model for multicategory purchase behavior of households. *J. Mark. Res.* **44** 595–612.

STERN, D. H., HERBRICH, R. and THORE, G. (2009). Matchbox: Large scale Bayesian recommendations. In 18*th International World Wide Web Conference*.

TITSIAS, M. K. (2016). One-vs-each approximation to softmax for scalable estimation of probabilities. In *Advances in Neural Information Processing Systems*.

TITSIAS, M. K. and LÁZARO-GREDILLA, M. (2014). Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*.

TRAIN, K. E., MCFADDEN, D. L. and BEN-AKIVA, M. (1987). The demand for local telephone service: A fully discrete model of residential calling patterns and service choices. *Rand J. Econ.* 109–123.

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

VILNIS, L. and MCCALLUM, A. (2015). Word representations via Gaussian embedding. In *International Conference on Learning Representations*.

WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

WAN, M., WANG, D., GOLDMAN, M., TADDY, M., RAO, J., LIU, J., LYMBEROPOULOS, D. and MCAULEY, J. (2017). Modeling consumer preferences and price sensitivities from large-scale grocery shopping transaction logs. In *International World Wide Web Conference*.

WANG, C. and BLEI, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the* 17*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

WOLPIN, K. I. (1984). An estimable dynamic stochastic model of fertility and child mortality. *J. Polit. Econ.* **92** 852–874.

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 28–50
https://doi.org/10.1214/19-AOAS1268

BART WITH TARGETED SMOOTHING: AN ANALYSIS OF PATIENT-SPECIFIC STILLBIRTH RISK

BY JENNIFER E. STARLING[1,*], JARED S. MURRAY[2,†], CARLOS M. CARVALHO[2,‡], RADEK K. BUKOWSKI[3] AND JAMES G. SCOTT[1,**]

[1]*Department of Statistics and Data Sciences, University of Texas at Austin,* [*]*jstarling@utexas.edu;* [**]*james.scott@mccombs.utexas.edu*

[2]*McCombs School of Business, University of Texas at Austin,* [†]*jared.murray@mccombs.utexas.edu;* [‡]*carlos.carvalho@mccombs.utexas.edu*

[3]*Department of Women's Health, Dell Medical School, University of Texas at Austin, radek.bukowski@austin.utexas.edu*

This article introduces BART with Targeted Smoothing, or tsBART, a new Bayesian tree-based model for nonparametric regression. The goal of tsBART is to introduce smoothness over a single target covariate t while not necessarily requiring smoothness over other covariates x. tsBART is based on the Bayesian Additive Regression Trees (BART) model, an ensemble of regression trees. tsBART extends BART by parameterizing each tree's terminal nodes with smooth functions of t rather than independent scalars. Like BART, tsBART captures complex nonlinear relationships and interactions among the predictors. But unlike BART, tsBART guarantees that the response surface will be smooth in the target covariate. This improves interpretability and helps to regularize the estimate.

After introducing and benchmarking the tsBART model, we apply it to our motivating example—pregnancy outcomes data from the National Center for Health Statistics. Our aim is to provide patient-specific estimates of stillbirth risk across gestational age (t) and based on maternal and fetal risk factors (x). Obstetricians expect stillbirth risk to vary smoothly over gestational age but not necessarily over other covariates, and tsBART has been designed precisely to reflect this structural knowledge. The results of our analysis show the clear superiority of the tsBART model for quantifying stillbirth risk, thereby providing patients and doctors with better information for managing the risk of fetal mortality. All methods described here are implemented in the R package *tsbart*.

REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

BAILIT, J. L., GREGORY, K. D., REDDY, U. M., GONZALEZ-QUINTERO, V. H., HIBBARD, J. U., RAMIREZ, M. M., BRANCH, D. W., BURKMAN, R., HABERMAN, S. et al. (2010). Maternal and neonatal outcomes by labor onset type and gestational age. *Am. J. Obstet. Gynecol.* **202** 245.e1–245.e12. https://doi.org/10.1016/j.ajog.2010.01.051

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93** 935–948.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172 https://doi.org/10.1214/09-AOAS285

CLARK, S. L., FRYE, D. R. and MYERS, J. A. (2010). Reduction in elective delivery at ≤ 39 weeks of gestation: Comparative effectiveness of 3 approaches to change and the impact on neonatal intensive care admission and stillbirth. *Am. J. Obstet. Gynecol.* **203** 449.e1–449.e6.

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

Key words and phrases. Bayesian additive regression tree, ensemble method, Gaussian process, regression tree, regularization.

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

GOLDSTEIN, A., KAPELNER, A., BLEICH, J. and PITKIN, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Statist*. **24** 44–65. MR3328247 https://doi.org/10.1080/10618600.2014.907095

GRAMACY, R. B. and LEE, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *J. Amer. Statist. Assoc*. **103** 1119–1130. MR2528830 https://doi.org/10.1198/016214508000000689

HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2017). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. Preprint. Available at arXiv:1706.09523v2.

HASTIE, T. and TIBSHIRANI, R. (2000). Bayesian backfitting. *Statist. Sci*. **15** 196–223. MR1820768 https://doi.org/10.1214/ss/1009212815

HE, J., SAAR, Y. and HAHN, P. R. (2018). Accelerated bayesian additive regression trees. Preprint. Available at arXiv:1810.02215.

HERNÁNDEZ, B., RAFTERY, A. E., PENNINGTON, S. R. and PARNELL, A. C. (2018). Bayesian additive regression trees using Bayesian model averaging. *Stat. Comput*. **28** 869–890. MR3766048 https://doi.org/10.1007/s11222-017-9767-1

HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist*. **20** 217–240. MR2816546 https://doi.org/10.1198/jcgs.2010.08162

KORNHAUSER, M. and SCHNEIDERMAN, R. (2010). How plans can improve outcomes and cut costs for preterm infant care. *Manag Care* **19** 28–30.

KRATZ, M. F. (2006). Level crossings and other level functionals of stationary Gaussian processes. *Probab. Surv*. **3** 230–288. MR2264709 https://doi.org/10.1214/154957806000000087

LINERO, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *J. Amer. Statist. Assoc*. **113** 626–636. MR3832214 https://doi.org/10.1080/01621459.2016.1264957

LINERO, A. R. and YANG, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **80** 1087–1110. MR3874311 https://doi.org/10.1111/rssb.12293

LOGAN, B. R., SPARAPANI, R., MCCULLOCH, R. E. and LAUD, P. W. (2019). Decision making and uncertainty quantification for individualized treatments using Bayesian additive regression trees. *Stat. Methods Med. Res*. **28** 1079–1093. MR3934636 https://doi.org/10.1177/0962280217746191

MACDORMAN, M. F. and GREGORY, E. C. W. (2015). Fetal and perinatal mortality: United States, 2013. *Natl. Vital Stat. Rep*. **66** 1–24.

MANDUJANO, A., WATERS, T. P. and MYERS, S. A. (2013). The risk of fetal death: Current concepts of best gestational age for delivery. *Am. J. Obstet. Gynecol*. **208** 207.e1–207.e8. https://doi.org/10.1016/j.ajog.2012.12.005

MURASKAS, J. and PARSI, K. (2008). The cost of saving the tiniest lives: NICUs versus prevention. *J. of Ethics* **10** 655–658.

MURRAY, J. S. (2017). Log-linear Bayesian additive regression trees for categorical and count responses. Preprint. Available at arXiv:1701.01503.

PRATOLA, M. T., CHIPMAN, H. A., GATTIKER, J. R., HIGDON, D. M., MCCULLOCH, R. and RUST, W. N. (2014). Parallel Bayesian additive regression trees. *J. Comput. Graph. Statist*. **23** 830–852. MR3224658 https://doi.org/10.1080/10618600.2013.841584

REDDY, U., BETTEGOWDA, V. R. and DIAS, T. (2011). Term pregnancy: A period of heterogeneous risk for infant mortality. *Obstet. Gynecol*. **117** 1279–1287.

SIVAGANESAN, S., MÜLLER, P. and HUANG, B. (2017). Subgroup finding via Bayesian additive regression trees. *Stat. Med*. **36** 2391–2403. MR3660139 https://doi.org/10.1002/sim.7276

SPARAPANI, R. A., LOGAN, B. R., MCCULLOCH, R. E. and LAUD, P. W. (2016). Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Stat. Med*. **35** 2741–2753. MR3513715 https://doi.org/10.1002/sim.6893

STARLING, J. E., MURRAY, J. S., CARVALHO, C. M., BUKOWSKI, R. K. and SCOTT, J. G. (2020). Supplement to "BART with targeted smoothing: An analysis of patient-specific stillbirth risk." https://doi.org/10.1214/19-AOAS1268SUPPA, https://doi.org/10.1214/19-AOAS1268SUPPB

WAGER, S., HASTIE, T. and EFRON, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *J. Mach. Learn. Res*. **15** 1625–1651. MR3225243

WATANABE, S. (2013). A widely applicable Bayesian information criterion. *J. Mach. Learn. Res*. **14** 867–897. MR3049492

XU, J., MURPHY, S. L., KOCHANEK, K. D. and BASTIAN, B. A. (2013). Deaths: Final data for 2013. *Natl. Vit. Stats. Rpt*. **64**.

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 51–73
https://doi.org/10.1214/19-AOAS1287

INTEGRATIVE SURVIVAL ANALYSIS WITH UNCERTAIN EVENT TIMES IN APPLICATION TO A SUICIDE RISK STUDY

BY WENJIE WANG[1,*], ROBERT ASELTINE[2], KUN CHEN[1,**] AND JUN YAN[1,†]

[1]*Department of Statistics, University of Connecticut,* [*]*wenjie.2.wang@uconn.edu;* [**]*kun.chen@uconn.edu;* [†]*jun.yan@uconn.edu*

[2]*Division of Behavioral Science and Community Health, Center for Population Health, UConn Health, aseltine@uchc.edu*

The concept of integrating data from disparate sources to accelerate scientific discovery has generated tremendous excitement in many fields. The potential benefits from data integration, however, may be compromised by the uncertainty due to incomplete/imperfect record linkage. Motivated by a suicide risk study, we propose an approach for analyzing survival data with uncertain event times arising from data integration. Specifically, in our problem deaths identified from the hospital discharge records together with reported suicidal deaths determined by the Office of Medical Examiner may still not include all the death events of patients, and the missing deaths can be recovered from a complete database of death records. Since the hospital discharge data can only be linked to the death record data by matching basic patient characteristics, a patient with a censored death time from the first dataset could be linked to multiple potential event records in the second dataset. We develop an integrative Cox proportional hazards regression in which the uncertainty in the matched event times is modeled probabilistically. The estimation procedure combines the ideas of profile likelihood and the expectation conditional maximization algorithm (ECM). Simulation studies demonstrate that under realistic settings of imperfect data linkage the proposed method outperforms several competing approaches including multiple imputation. A marginal screening analysis using the proposed integrative Cox model is performed to identify risk factors associated with death following suicide-related hospitalization in Connecticut. The identified diagnostics codes are consistent with existing literature and provide several new insights on suicide risk, prediction and prevention.

REFERENCES

ALBERT, A. and ANDERSON, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10. MR0738319 https://doi.org/10.1093/biomet/71.1.1

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

BOHENSKY, M. A., JOLLEY, D., SUNDARARAJAN, V., EVANS, S., PILCHER, D. V., SCOTT, I. and BRAND, C. A. (2010). Data linkage: A powerful research tool with potential problems. *BMC Health Serv. Res.* **10** 346. https://doi.org/10.1186/1472-6963-10-346.

BOSTWICK, M. J., PABBATI, C., GESKE, J. R. and MCKEAN, A. J. (2015). Suicide attempt as a risk factor for completed suicide: Even more lethal than we knew. *Am. J. Psychiatr.* **173** 1094–1100.

BRESLOW, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30** 89–99.

CHEN, K. and ASELTINE, R. (2017). Using hospitalization and mortality data to target suicide prevention activities: A demonstration from Connecticut. *J. Adolesc. Health* **61** 192–197.

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

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276. MR0400509 https://doi.org/10.1093/biomet/62.2.269

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

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

Key words and phrases. Cox model, data linkage, ECM algorithm, integrative learning, suicide prevention.

EFRON, B. (1981). Censored data and the bootstrap. *J. Amer. Statist. Assoc.* **76** 312–319. MR0624333
FLEET, R. P., DUPUIS, G., MARCHAND, A., BURELLE, D., ARSENAULT, A. and BEITMAN, B. D. (1996). Panic disorder in emergency department chest pain patients: Prevalence, comorbidity, suicidal ideation, and physician recognition. *Am. J. Med.* **101** 371–380.
HADI, A. S. and LUCEÑO, A. (1997). Maximum trimmed likelihood estimators: A unified approach, examples, and algorithms. *Comput. Statist. Data Anal.* **25** 251–272. MR1478539 https://doi.org/10.1016/S0167-9473(97)00011-X
HARRIS, E. C. and BARRACLOUGH, B. (1997). Suicide as an outcome for mental disorders. A meta-analysis. *Br. J. Psychiatry* **170** 205–228.
HARRON, K., GOLDSTEIN, H. and DIBBEN, C. (2015). *Methodological Developments in Data Linkage*. Wiley.
HEAGERTY, P. J. and ZHENG, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61** 92–105. MR2135849 https://doi.org/10.1111/j.0006-341X.2005.030814.x
HOF, M. H. P. and ZWINDERMAN, A. H. (2012). Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Stat. Med.* **31** 4231–4242. MR3040077 https://doi.org/10.1002/sim.5498
HOF, M. H. P. and ZWINDERMAN, A. H. (2015). A mixture model for the analysis of data derived from record linkage. *Stat. Med.* **34** 74–92. MR3286240 https://doi.org/10.1002/sim.6315
JAMSHIDIAN, M. and JENNRICH, R. I. (2000). Standard errors for EM estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 257–270. MR1749538 https://doi.org/10.1111/1467-9868.00230
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
KATON, W., HALL, M. L., RUSSO, J., CORMIER, L., HOLLIFIELD, M., VITALIANO, P. P. and BEITMAN, B. D. (1988). Chest pain: Relationship of psychiatric illness to coronary arteriographic results. *Am. J. Med.* **84** 1–9. https://doi.org/10.1016/0002-9343(88)90001-0
KOPONEN, H., KAUTIAINEN, H., LEPPÄNEN, E., MÄNTYSELKÄ, P. and VANHALA, M. (2015). Association between suicidal behaviour and impaired glucose metabolism in depressive disorders. *BMC Psychiatry* **15** 163. https://doi.org/10.1186/s12888-015-0567-x
KUNG, H.-C., PEARSON, J. L. and WEI, R. (2005). Substance use, firearm availability, depressive symptoms, and mental health service utilization among White and African American suicide decedents aged 15 to 64 years. *Ann. Epidemiol.* **15** 614–621.
LIEB, K., ZANARINI, M. C., SCHMAHL, C., LINEHAN, M. M. and BOHUS, M. (2004). Borderline personality disorder. *Lancet* **364** 453–461.
MCGIRR, A., PARIS, J., LESAGE, A., RENAUD, J. and TURECKI, G. (2007). Risk factors for suicide completion in borderline personality disorder: A case-control study of cluster B comorbidity and impulsive aggression. *J. Clin. Psychiatry* **68** 721–729.
MEIER, A. S., RICHARDSON, B. A. and HUGHES, J. P. (2003). Discrete proportional hazards models for mismeasured outcomes. *Biometrics* **59** 947–954. MR2025118 https://doi.org/10.1111/j.0006-341X.2003.00109.x
MENG, X.-L. and RUBIN, D. B. (1991). Using EM to obtain asymptotic variance–covariance matrices: The SEM algorithm. *J. Amer. Statist. Assoc.* **86** 899–909.
MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503 https://doi.org/10.1093/biomet/80.2.267
MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. MR1803168 https://doi.org/10.2307/2669386
NADARAJAH, S. and KOTZ, S. (2006). R programs for truncated distributions. *J. Stat. Softw.* **16** 1–8.
NEYKOV, N., FILZMOSER, P., DIMOVA, R. and NEYTCHEV, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Comput. Statist. Data Anal.* **52** 299–308. MR2409983 https://doi.org/10.1016/j.csda.2006.12.024
PATRICK, A. R., MILLER, M., BARBER, C. W., WANG, P. S., CANNING, C. F. and SCHNEEWEISS, S. (2010). Identification of hospitalizations for intentional self-harm when E-codes are incompletely recorded. *Pharmacoepidemiol. Drug Saf.* **19** 1263–1275. https://doi.org/10.1002/pds.2037.
PENA, J. B., MATTHIEU, M. M., ZAYAS, L. H., MASYN, K. E. and CAINE, E. D. (2012). Co-occurring risk behaviors among White, Black, and Hispanic US high school adolescents with suicide attempts requiring medical attention, 1999–2007: Implications for future prevention initiatives. *Soc. Psychiatry Psychiatr. Epidemiol.* **47** 29–42. https://doi.org/10.1007/s00127-010-0322-z.
PRITCHARD, C. and HANSEN, L. (2015). Examining undetermined and accidental deaths as source of 'under-reported-suicide' by age and sex in twenty western countries. *Community Ment. Health J.* **51** 365–376. https://doi.org/10.1007/s10597-014-9810-z.
R DEVELOPMENT CORE TEAM (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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.

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

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519 https://doi.org/10.1002/9780470316696

SNAPINN, S. M. (1998). Survival analysis with uncertain endpoints. *Biometrics* **54** 209–218.

SUOMINEN, K., ISOMETSÄ, E., SUOKAS, J., HAUKKA, J., ACHTE, K. and LÖNNQVIST, J. (2004). Completed suicide after a suicide attempt: A 37-year follow-up study. *Am. J. Psychiatr.* **161** 562–563.

TANCREDI, A. and LISEO, B. (2015). Regression analysis with linked data: Problems and possible solutions. *Statistica* **75** 19–35.

TØLLEFSEN, I. M., THIBLIN, I., HELWEG-LARSEN, K., HEM, E., KASTRUP, M., NYBERG, U., ROGDE, S., ZAHL, P.-H., ØSTEVOLD, G. et al. (2016). Accidents and undetermined deaths: Re-evaluation of nationwide samples from the Scandinavian countries. *BMC Public Health* **16** 449. https://doi.org/10.1186/s12889-016-3135-5.

WANG, W., ASELTINE, R., CHEN, K. and YAN, J. (2020). Supplement to "Integrative survival analysis with uncertain event times in application to a suicide risk study." https://doi.org/10.1214/19-AOAS1287SUPP.

WINGLEE, M., VALLIANT, R. and SCHEUREN, F. (2005). A case study in record linkage. *Surv. Methodol.* **31** 3–11.

XU, C., BAINES, P. D. and WANG, J.-L. (2014). Standard error estimation using the EM algorithm for the joint modeling of survival and longitudinal data. *Biostatistics* **15** 731–744. https://doi.org/10.1093/biostatistics/kxu015.

ZHAO, Q., SHI, X., XIE, Y., HUANG, J., SHIA, B. and MA, S. (2015). Combining multidimensional genomic measurements for predicting cancer prognosis: Observations from TCGA. *Brief. Bioinform.* **16** 291–303. https://doi.org/10.1093/bib/bbu003.

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 74–93
https://doi.org/10.1214/19-AOAS1278

EFFICIENT REAL-TIME MONITORING OF AN EMERGING INFLUENZA PANDEMIC: HOW FEASIBLE?

BY PAUL J. BIRRELL[1,*], LORENZ WERNISCH[1,**], BRIAN D. M. TOM[1,†], LEONHARD HELD[2], GARETH O. ROBERTS[3], RICHARD G. PEBODY[4] AND DANIELA DE ANGELIS[1,‡]

[1]*MRC Biostatistics Unit, Cambridge Institute of Public Health, University of Cambridge,* [*]*paul.birrell@mrc-bsu.cam.ac.uk;* [**]*lorenz.wernisch@mrc-bsu.cam.ac.uk;* [†]*brian.tom@mrc-bsu.cam.ac.uk;* [‡]*daniela.deangelis@mrc-bsu.cam.ac.uk*

[2]*Epidemiology, Biostatistics and Prevention Institute (EBPI), University of Zurich, leonhard.held@uzh.ch*

[3]*CRiSM, Department of Statistics, University of Warwick, Gareth.O.Roberts@warwick.ac.uk*

[4]*Health Protection Directorate, Public Health England, richard.pebody@phe.gov.uk*

A prompt public health response to a new epidemic relies on the ability to monitor and predict its evolution in real time as data accumulate. The 2009 A/H1N1 outbreak in the UK revealed pandemic data as noisy, contaminated, potentially biased and originating from multiple sources. This seriously challenges the capacity for real-time monitoring. Here, we assess the feasibility of real-time inference based on such data by constructing an analytic tool combining an age-stratified SEIR transmission model with various observation models describing the data generation mechanisms. As batches of data become available, a sequential Monte Carlo (SMC) algorithm is developed to synthesise multiple imperfect data streams, iterate epidemic inferences and assess model adequacy amidst a rapidly evolving epidemic environment, substantially reducing computation time in comparison to standard MCMC, to ensure timely delivery of real-time epidemic assessments. In application to simulated data designed to mimic the 2009 A/H1N1 epidemic, SMC is shown to have additional benefits in terms of assessing predictive performance and coping with parameter nonidentifiability.

REFERENCES

AHRENS, H. (1976). Multivariate variance-covariance components (MVCC) and generalized intraclass correlation coefficient (GICC). *Biom. J.* **18** 527–533.

BANTERLE, M., GRAZIAN, C., LEE, A. and ROBERT, C. P. (2019). Accelerating Metropolis–Hastings algorithms by Delayed Acceptance. *Foundations of Data Science* **1** 103–128.

BETTENCOURT, L. M. A. and RIBEIRO, R. M. (2008). Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS ONE* **3** e2185. https://doi.org/10.1371/journal.pone.0002185

BIERKENS, J., FEARNHEAD, P. and ROBERTS, G. (2019). The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data. *Ann. Stat.* **47** 1288–1320.

BIRRELL, P. J., KETSETZIS, G., GAY, N. G., COOPER, B. S., PRESANIS, A. M., HARRIS, R. J., CHARLETT, A., ZHANG, X.-S., WHITE, P. et al. (2011). Bayesian modelling to unmask and predict the influenza A/H1N1pdm dynamics in London. *Proc. Natn. Acad. Sci. USA* **108** 18238–18243.

BIRRELL, P. J., WERNISCH, L., TOM, B. D. M., HELD, L., ROBERTS, G. O., PEBODY, R. G. and DE ANGELIS, D. (2020). Supplement to "Efficient real-time monitoring of an emerging influenza pandemic: How feasible?." https://doi.org/10.1214/19-AOAS1278SUPP.

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

CAMACHO, A., KUCHARSKI, A., AKI-SAWYERR, Y., WHITE, M. A., FLASCHE, S., BAGUELIN, M., POLLINGTON, T., CARNEY, J. R., GLOVER, R. et al. (2015). Temporal changes in Ebola transmission in Sierra Leone and implications for control requirements: A real-time modelling study. *PLoS Curr* **7**.

CARPENTER, J., CLIFFORD, P. and FEARNHEAD, P. (1999). Improved particle filter for nonlinear problems. *IEE Proc. Radar Sonar Navig.* **146** 2+.

Key words and phrases. Sequential Monte Carlo, resample-move, real-time inference, pandemic influenza, SEIR transmission model.

CAUCHEMEZ, S., BOËLLE, P. Y., THOMAS, G. and VALLERON, A. J. (2006). Estimating in real time the efficacy of measures to control emerging communicable diseases. *Am. J. Epidemiol.* **164** 591–597.

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

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

DAWID, A. P. (1984). Statistical theory. The prequential approach. *J. Roy. Statist. Soc. Ser. A* **147** 278–292. MR0763811 https://doi.org/10.2307/2981683

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

DONNER, A. and KOVAL, J. J. (1980). The estimation of intraclass correlation in the analysis of family data. *Biometrics* **36** 19–25.

DOUCET, A. and JOHANSEN, A. M. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In *The Oxford Handbook of Nonlinear Filtering* 656–704. Oxford Univ. Press, Oxford. MR2884612

DUKIC, V., LOPES, H. F. and POLSON, N. G. (2012). Tracking epidemics with Google Flu Trends data and a state-space SEIR model. *J. Amer. Statist. Assoc.* **107** 1410–1426. MR3036404 https://doi.org/10.1080/01621459.2012.713876

DUREAU, J., KALOGEROPOULOS, K. and BAGUELIN, M. (2013). Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. *Biostatistics* **14** 541–555.

FARAH, M., BIRRELL, P., CONTI, S. and DE ANGELIS, D. (2014). Bayesian emulation and calibration of a dynamic epidemic model for A/H1N1 influenza. *J. Amer. Statist. Assoc.* **109** 1398–1411. MR3293599 https://doi.org/10.1080/01621459.2014.934453

FEARNHEAD, P. (2002). Markov chain Monte Carlo, sufficient statistics, and particle filters. *J. Comput. Graph. Statist.* **11** 848–862. MR1951601 https://doi.org/10.1198/106186002321018821

FEARNHEAD, P. and TAYLOR, B. M. (2013). An adaptive sequential Monte Carlo sampler. *Bayesian Anal.* **8** 411–438. MR3066947 https://doi.org/10.1214/13-BA814

FUNK, S., CAMACHO, A., KUCHARSKI, A. J., EGGO, R. M. and EDMUNDS, W. J. (2018). Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. *Epidemics* **22** 56–61. https://doi.org/10.1016/j.epidem.2016.11.003

GEYER, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics*: *The 23rd Symposium on the Interface* 156–163. Interface Foundation of North America, Fairfax Station, VA.

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

GIROLAMI, M. and CALDERHEAD, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 123–214. MR2814492 https://doi.org/10.1111/j.1467-9868.2010.00765.x

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548 https://doi.org/10.1198/016214506000001437

GORDON, N. J., SALMOND, D. J. and SMITH, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc-F* **140** 107–113.

HELD, L., MEYER, S. and BRACHER, J. (2017). Probabilistic forecasting in infectious disease epidemiology: The 13th Armitage lecture. *Stat. Med.* **36** 3443–3460. MR3696502 https://doi.org/10.1002/sim.7363

JASRA, A., STEPHENS, D. A. and HOLMES, C. C. (2007). On population-based simulation for static inference. *Stat. Comput.* **17** 263–279. MR2405807 https://doi.org/10.1007/s11222-007-9028-9

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

JEWELL, C. P., KYPRAIOS, T., CHRISTLEY, R. M. and ROBERTS, G. O. (2009). A novel approach to real-time risk prediction for emerging infectious diseases: A case study in Avian Influenza H5N1. *Prev. vet. med.* **91** 19–28.

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

KONISHI, S., KHATRI, C. G. and RAO, C. R. (1991). Inferences on multivariate measures of interclass and intraclass correlations in familial data. *J. Roy. Statist. Soc. Ser. B* **53** 649–659. MR1125722

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

LIU, J. S. and CHEN, R. (1995). Blind deconvolution via sequential imputations. *J. Amer. Statist. Assoc.* **90** 567–576. MR3363399 https://doi.org/10.1080/01621459.1995.10476549

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

MEESTER, R., DE KONING, J., DE JONG, M. C. M. and DIEKMANN, O. (2002). Modeling and real-time prediction of classical swine fever epidemics. *Biometrics* **58** 178–184. MR1891377 https://doi.org/10.1111/j.0006-341X.2002.00178.x

NEAL, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Stat. Comput.* **6** 353–366.

NEMETH, C., FEARNHEAD, P. and MIHAYLOVA, L. (2014). Sequential Monte Carlo methods for state and parameter estimation in abruptly changing environments. *IEEE Trans. Signal Process.* **62** 1245–1255. MR3168149 https://doi.org/10.1109/TSP.2013.2296278

ONG, J. B. S., CHEN, M. I.-C., COOK, A. R., CHYI, H., LEE, V. J., PIN, R. T., ANANTH, P. and GAN, L. (2010). Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore. *PLoS ONE* **5** e10036.

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

SCHUSTER, I., STRATHMANN, H., PAIGE, B. and SEJDINOVIC, D. (2017). Kernel sequential Monte Carlo. In *Machine Learning and Knowledge Discovery in Databases* (M. Ceci, J. Hollmén, L. Todorovski, C. Vens and S. Džeroski, eds.) 390–409. Springer, Cham.

SCIENTIFIC PANDEMIC INFLUENZA ADVISORY COMMITTEE: SUBGROUP ON MODELLING (2011). Modelling Summary. SPI-M-O Committee document (Accessed 4 February, 2016).

SEILLIER-MOISEIWITSCH, F. and DAWID, A. P. (1993). On testing the validity of sequential probability forecasts. *J. Amer. Statist. Assoc.* **88** 355–359. MR1212496

SHAMAN, J. and KARSPECK, A. (2012). Forecasting seasonal outbreaks of influenza. *Proc. Natn. Acad. Sci. USA* **109** 20425–20430.

SHERLOCK, C., FEARNHEAD, P. and ROBERTS, G. O. (2010). The random walk Metropolis: Linking theory and practice through a case study. *Statist. Sci.* **25** 172–190. MR2789988 https://doi.org/10.1214/10-STS327

SHUBIN, M., LEBEDEV, A., LYYTIKÄINEN, O. and AURANEN, K. (2016). Revealing the true incidence of pandemic A(H1N1)pdm09 influenza in Finland during the first two seasons—an analysis based on a dynamic transmission model. *PLoS Comput. Biol.* **12** 1–3.

SKVORTSOV, A. and RISTIC, B. (2012). Monitoring and prediction of an epidemic outbreak using syndromic observations. *Math. Biosci.* **240** 12–19. MR2974537 https://doi.org/10.1016/j.mbs.2012.05.010

SOKAL, R. R. and ROHLF, F. (1981). *Biometry*, 2nd ed. **668**. WH Feeman and Company, New York.

TE BEEST, D. E., BIRRELL, P. J., WALLINGA, J., ANGELIS, D. D. and VAN BOVEN, M. (2015). Joint modelling of serological and hospitalization data reveals that high levels of pre-existing immunity and school holidays shaped the influenza A pandemic of 2009 in the Netherlands. *J. R. Soc. Interface* **12**. https://doi.org/10.1098/rsif.2014.1244

VIBOUD, C., SUN, K., GAFFEY, R., AJELLI, M., FUMANELLI, L., MERLER, S., ZHANG, Q., CHOWELL, G., SIMONSEN, L. et al. (2018). The RAPIDD Ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics* **22** 13–21.

WALLINGA, J. and TEUNIS, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* **160** 509–516.

WEST, M. (1993). Mixtures models, Monte Carlo, Bayesian updating and dynamic models. *Computer Science and Statistics* **24** 325–333.

WHITELEY, N., JOHANSEN, A. M. and GODSILL, S. (2011). Monte Carlo filtering of piecewise deterministic processes. *J. Comput. Graph. Statist.* **20** 119–139. MR2816541 https://doi.org/10.1198/jcgs.2009.08052

WU, J. T., COWLING, B. J., LAU, E. H. Y., IP, D. K. M., HO, L. M., TSANG, T., CHUANG, S. K., LEUNG, P. Y., LO, S. V. et al. (2010). School closure and mitigation of pandemic (H1N1) 2009, Hong Kong. *Emerg. Infec. Dis.* **16** 538–541.

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 94–115
https://doi.org/10.1214/19-AOAS1283

MODELING MICROBIAL ABUNDANCES AND DYSBIOSIS WITH BETA-BINOMIAL REGRESSION

BY BRYAN D. MARTIN[1], DANIELA WITTEN[2] AND AMY D. WILLIS[3]

[1]*Department of Statistics, University of Washington, bmartin6@uw.edu*
[2]*Departments of Statistics and Biostatistics, University of Washington, dwitten@uw.edu*
[3]*Department of Biostatistics, University of Washington, adwillis@uw.edu*

Using a sample from a population to estimate the proportion of the population with a certain category label is a broadly important problem. In the context of microbiome studies, this problem arises when researchers wish to use a sample from a population of microbes to estimate the population proportion of a particular taxon, known as the taxon's *relative abundance*. In this paper, we propose a beta-binomial model for this task. Like existing models, our model allows for a taxon's relative abundance to be associated with covariates of interest. However, unlike existing models, our proposal also allows for the overdispersion in the taxon's counts to be associated with covariates of interest. We exploit this model in order to propose tests not only for differential relative abundance, but also for differential variability. The latter is particularly valuable in light of speculation that *dysbiosis*, the perturbation from a normal microbiome that can occur in certain disease conditions, may manifest as a loss of stability, or increase in variability, of the counts associated with each taxon. We demonstrate the performance of our proposed model using a simulation study and an application to soil microbial data.

REFERENCES

AERTS, M., MOLENBERGHS, G., GEYS, H. and RYAN, L. M. (2002). *Topics in Modelling of Clustered Data*. CRC Press/CRC, Boca Raton, FL.

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability*. CRC Press, London. MR0865647 https://doi.org/10.1007/978-94-009-4109-0

ALBERT, A. and ANDERSON, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10. MR0738319 https://doi.org/10.1093/biomet/71.1.1

BASTEDO, M. N. and JAQUETTE, O. (2011). Running in place: Low-income students and the dynamics of higher education stratification. *Educ. Eval. Policy Anal.* **33** 318–339.

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

CALLAHAN, B. J., DIGIULIO, D. B., GOLTSMAN, D. S. A., SUN, C. L., COSTELLO, E. K., JEGANATHAN, P., BIGGIO, J. R., WONG, R. J., DRUZIN, M. L. et al. (2017). Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women. *Proc. Natl. Acad. Sci. USA* **114** 9966–9971.

CAO, Y., ZHANG, A. and LI, H. (2017). Microbial composition estimation from sparse count data. Preprint. Available at arXiv:1706.02380.

CHAI, H., JIANG, H., LIN, L. and LIU, L. (2018). A marginalized two-part Beta regression model for microbiome compositional data. *PLoS Comput. Biol.* **14** e1006329.

CHEN, J. and LI, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* **7** 418–442. MR3086425 https://doi.org/10.1214/12-AOAS592

CHEN, E. Z. and LI, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* **32** 2611–2617.

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

Key words and phrases. Relative abundance, microbiome, correlated data, overdispersion, high throughput sequencing, beta-binomial.

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

DIGIULIO, D. B., CALLAHAN, B. J., MCMURDIE, P. J., COSTELLO, E. K., LYELL, D. J., ROBACZEWSKA, A., SUN, C. L., GOLTSMAN, D. S. A., WONG, R. J. et al. (2015). Temporal and spatial variation of the human microbiota during pregnancy. *Proc. Natl. Acad. Sci. USA* **112** 11060–11065.

DOLZHENKO, E. and SMITH, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinform.* **15** 215. https://doi.org/10.1186/1471-2105-15-215

EDGAR, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* **10** 996–998. https://doi.org/10.1038/nmeth.2604

FANG, R., WAGNER, B. D., HARRIS, J. K. and FILLON, S. A. (2016). Zero-inflated negative binomial mixed model: An application to two microbial organisms important in oesophagitis. *Epidemiol. Infect.* **144** 2447–2455.

FAUST, K., LAHTI, L., GONZE, D., DE VOS, W. M. and RAES, J. (2015). Metagenomics meets time series analysis: Unraveling microbial community dynamics. *Curr. Opin. Microbiol.* **25** 56–66. https://doi.org/10.1016/j.mib.2015.04.004

FIACCO, A. V. and MCCORMICK, G. P. (1968). *Nonlinear Programming*: *Sequential Unconstrained Minimization Techniques*. Wiley, New York. MR0243831

FLETCHER, R. (1987). *Practical Methods of Optimization*, 2nd ed. Wiley, Chichester. MR0955799

GERBER, G. K. (2014). The dynamic microbiome. *FEBS Lett.* **588** 4131–4139.

GEVERS, D., KUGATHASAN, S., DENSON, L. A., VÁZQUEZ-BAEZA, Y., VAN TREUREN, W., REN, B., SCHWAGER, E., KNIGHTS, D., SONG, S. J. et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15** 382–392.

GEYER, C. J. (2015). trust: Trust region optimization. R package version 0.1-7.

GRICE, E. A. (2014). The skin microbiome: Potential for novel diagnostic and therapeutic approaches to cutaneous disease. *Semin. Cutan. Med. Surg.* **33** 98. NIH Public Access.

HALFVARSON, J., BRISLAWN, C. J., LAMENDELLA, R., VÁZQUEZ-BAEZA, Y., WALTERS, W. A., BRAMER, L. M., D'AMATO, M., BONFIGLIO, F., MCDONALD, D. et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.* **2** 17004. https://doi.org/10.1038/nmicrobiol.2017.4

HEINZE, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat. Med.* **25** 4216–4226. MR2307586 https://doi.org/10.1002/sim.2687

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

HILL-BURNS, E. M., DEBELIUS, J. W., MORTON, J. T., WISSEMANN, W. T., LEWIS, M. R., WALLEN, Z. D., PEDDADA, S. D., FACTOR, S. A., MOLHO, E. et al. (2017). Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. *Mov. Disord.* **32** 739–749. https://doi.org/10.1002/mds.26942

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

HOOKS, K. B. and O'MALLEY, M. A. (2017). Dysbiosis and its discontents. *mBio* **8** e01492-17. https://doi.org/10.1128/mBio.01492-17

KLEINMAN, J. C. (1973). Proportions with extraneous variance: Single and independent samples. *J. Amer. Statist. Assoc.* **68** 46–54.

KOSMIDIS, I. (2018). brglm2: Bias reduction in generalized linear models. R package version 0.1.8.

KURTZ, Z. D., MÜLLER, C. L., MIRALDI, E. R., LITTMAN, D. R., BLASER, M. J. and BONNEAU, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11** e1004226. https://doi.org/10.1371/journal.pcbi.1004226

LAW, C. W., CHEN, Y., SHI, W. and SMYTH, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15** R29.

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

LI, Z., LEE, K., KARAGAS, M. R., MADAN, J. C., HOEN, A. G., O'MALLEY, A. J. and LI, H. (2018). Conditional regression based on a multivariate zero-inflated logistic-normal model for microbiome relative abundance data. *Stat. Biosci.* **10** 587–608.

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

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

MARTIN, B. D., WITTEN, D. and WILLIS, A. D. (2020a). Supplement A to "Modeling microbial abundances and dysbiosis with beta-binomial regression." https://doi.org/10.1214/19-AOAS1283SUPPA.

MARTIN, B. D., WITTEN, D. and WILLIS, A. D. (2020b). Supplement B to "Modeling microbial abundances and dysbiosis with beta-binomial regression." https://doi.org/10.1214/19-AOAS1283SUPPB.

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

MCMURDIE, P. J. and HOLMES, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8** e61217.

MCMURDIE, P. J. and HOLMES, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Comput*. *Biol*. **10** e1003531.

MERCER, L. D., WAKEFIELD, J., PANTAZIS, A., LUTAMBI, A. M., MASANJA, H. and CLARK, S. (2015). Space-time smoothing of complex survey data: Small area estimation for child mortality. *Ann*. *Appl*. *Stat*. **9** 1889–1905. MR3456357 https://doi.org/10.1214/15-AOAS872

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

MORGAN, X. C., KABAKCHIEV, B., WALDRON, L., TYLER, A. D., TICKLE, T. L., MILGROM, R., STEMPAK, J. M., GEVERS, D., XAVIER, R. J. et al. (2015). Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biol*. **16** 67. https://doi.org/10.1186/s13059-015-0637-x

NOCEDAL, J. and WRIGHT, S. J. (1999). *Numerical Optimization*. *Springer Series in Operations Research*. Springer, New York. MR1713114 https://doi.org/10.1007/b98874

PARKER, I. M., SAUNDERS, M., BONTRAGER, M., WEITZ, A. P., HENDRICKS, R., MAGAREY, R., SUITER, K. and GILBERT, G. S. (2015). Phylogenetic structure and host abundance drive disease pressure in communities. *Nature* **520** 542–544. https://doi.org/10.1038/nature14372

PAULSON, J. N., STINE, O. C., BRAVO, H. C. and POP, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat*. *Methods* **10** 1200–1202.

PENG, X., LI, G. and LIU, Z. (2016). Zero-inflated beta regression for differential abundance analysis with metagenomics data. *J*. *Comput*. *Biol*. **23** 102–110.

PETERSEN, C. and ROUND, J. L. (2014). Defining dysbiosis and its influence on host immunity and disease. *Cell*. *Microbiol*. **16** 1024–1033.

POUSSIN, C., SIERRO, N., BOUÉ, S., BATTEY, J., SCOTTI, E., BELCASTRO, V., PEITSCH, M. C., IVANOV, N. V. and HOENG, J. (2018). Interrogating the microbiome: Experimental and computational considerations in support of study reproducibility. *Drug Discov. Today* **23** 1644–1657. https://doi.org/10.1016/j.drudis.2018.06.005

PRENTICE, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J*. *Amer*. *Statist*. *Assoc*. **81** 321–327.

QIN, N., YANG, F., LI, A., PRIFTI, E., CHEN, Y., SHAO, L., GUO, J., LE CHATELIER, E., YAO, J. et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513** 59.

R CORE TEAM (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

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

ROBINSON, M. D. and OSHLACK, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. **11** R25. https://doi.org/10.1186/gb-2010-11-3-r25

RYAN, D. M. (1974). Penalty and barrier functions. In *Numerical Methods for Constrained Optimization* (*Proc*. *Sympos*., *National Physical Lab*., *Teddington*, 1974) 175–190. MR0456505

SANKARAN, K. and HOLMES, S. P. (2017). Latent variable modeling for the microbiome. Preprint. Available at arXiv:1706.04969.

SEGATA, N., IZARD, J., WALDRON, L., GEVERS, D., MIROPOLSKY, L., GARRETT, W. S. and HUTTENHOWER, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol*. **12** R60. https://doi.org/10.1186/gb-2011-12-6-r60

SENDER, R., FUCHS, S. and MILO, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol*. **14** e1002533. https://doi.org/10.1371/journal.pbio.1002533

SHI, B., CHANG, M., MARTIN, J., MITREVA, M., LUX, R., KLOKKEVOLD, P., SODERGREN, E., WEINSTOCK, G. M., HAAKE, S. K. et al. (2015). Dynamic changes in the subgingival microbiome and their potential for diagnosis and prognosis of periodontitis. *mBio* **6** e01926-14.

SKELLAM, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **10** 257–261. MR0028539

SOGIN, M. L., MORRISON, H. G., HUBER, J. A., WELCH, D. M., HUSE, S. M., NEAL, P. R., ARRIETA, J. M. and HERNDL, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proc. Natl. Acad. Sci. USA* **103** 12115–12120.

SOHN, M. B., DU, R. and AN, L. (2015). A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics* **31** 2269–2275.

TAMBOLI, C. P., NEUT, C., DESREUMAUX, P. and COLOMBEL, J. F. (2004). Dysbiosis in inflammatory bowel disease. *Gut* **53** 1–4. https://doi.org/10.1136/gut.53.1.1

TROMAS, N., TARANU, Z. E., MARTIN, B. D., WILLIS, A., FORTIN, N., GREER, C. W. and SHAPIRO, B. J. (2018). Niche separation increases with genetic distance among bloom-forming cyanobacteria. *Front. Microbiol.* **9** 438. https://doi.org/10.3389/fmicb.2018.00438

WAGNER, B., RIGGS, P. and MIKULICH-GILBERTSON, S. (2015). The importance of distribution-choice in modeling substance use data: A comparison of negative binomial, beta binomial, and zero-inflated distributions. *Am. J. Drug Alcohol Abuse* **41** 489–497.

WAHBA, G., WANG, Y., GU, C., KLEIN, R. and KLEIN, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.* **23** 1865–1895. MR1389856 https://doi.org/10.1214/aos/1034713638

WELCH, J. L. M., ROSSETTI, B. J., RIEKEN, C. W., DEWHIRST, F. E. and BORISY, G. G. (2016). Biogeography of a human oral microbiome at the micron scale. *Proc. Natl. Acad. Sci. USA* **113** E791–E800.

WHITE, J. R., NAGARAJAN, N. and POP, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* **5** e1000352. https://doi.org/10.1371/journal.pcbi.1000352

WHITMAN, T., PEPE-RANNEY, C., ENDERS, A., KOECHLI, C., CAMPBELL, A., BUCKLEY, D. H. and LEHMANN, J. (2016). Dynamics of microbial community composition and soil organic carbon mineralization in soil following addition of pyrogenic and fresh organic matter. *ISME J.* **10** 2918–2930. https://doi.org/10.1038/ismej.2016.68

WICKHAM, H. (2016). *ggplot2*: *Elegant Graphics for Data Analysis*. Springer, New York.

WILLIAMS, D. A. (1975). 394: The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31** 949–952.

WILLIS, A. D. and MARTIN, B. D. (2018). DivNet: Estimating diversity in networked communities. *BioRxiv* 305045.

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

YEE, T. W. (2010). The VGAM package for categorical data analysis. *J. Stat. Softw.* **32** 1–34.

ZHANG, X., MALLICK, H., TANG, Z., ZHANG, L., CUI, X., BENSON, A. K. and YI, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinform.* **18** 4.

ZHOU, Y., SHAN, G., SODERGREN, E., WEINSTOCK, G., WALKER, W. A. and GREGORY, K. E. (2015). Longitudinal analysis of the premature infant intestinal microbiome prior to necrotizing enterocolitis: A case-control study. *PLoS ONE* **10** e0118632. https://doi.org/10.1371/journal.pone.0118632

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 116–142
https://doi.org/10.1214/19-AOAS1290

A STATISTICAL ANALYSIS OF NOISY CROWDSOURCED WEATHER DATA

BY ARNAB CHAKRABORTY*, SOUMENDRA NATH LAHIRI AND ALYSON WILSON

*Department of Statistics, North Carolina State University, * arnab2897@gmail.com*

Spatial prediction of weather elements like temperature, precipitation, and barometric pressure are generally based on satellite imagery or data collected at ground stations. None of these data provide information at a more granular or "hyperlocal" resolution. On the other hand, crowdsourced weather data, which are captured by sensors installed on mobile devices and gathered by weather-related mobile apps like `WeatherSignal` and `AccuWeather`, can serve as potential data sources for analyzing environmental processes at a hyperlocal resolution. However, due to the low quality of the sensors and the nonlaboratory environment, the quality of the observations in crowdsourced data is compromised. This paper describes methods to improve hyperlocal spatial prediction using this varying-quality, noisy crowdsourced information. We introduce a reliability metric, namely Veracity Score (VS), to assess the quality of the crowdsourced observations using a coarser, but high-quality, reference data. A VS-based methodology to analyze noisy spatial data is proposed and evaluated through extensive simulations. The merits of the proposed approach are illustrated through case studies analyzing crowdsourced daily average ambient temperature readings for one day in the contiguous United States.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th ed. Dover, New York.

ACCUWEATHER (2015). AccuWeather launches AccUcast, providing exclusive crowdsourced weather feature worldwide. Available at https://www.accuweather.com/en/press/50601069. Accessed: 2019-01-30.

ALLAHBAKHSH, M., BENATALLAH, B., IGNJATOVIC, A., MOTAHARI-NEZHAD, H. R., BERTINO, E. and DUSTDAR, S. (2013). Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Comput.* **17** 76–81.

CHAKRABORTY, A. and LAHIRI, S. N. (2019). On statistical properties of a veracity scoring method for spatial data. arXiv preprint arXiv:1906.08843.

CHAKRABORTY, A., LAHIRI, S. N. and WILSON, A. (2020). Supplement to "A statistical analysis of noisy crowdsourced weather data." https://doi.org/10.1214/19-AOAS1290SUPP.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York.

CRESSIE, N. and HAWKINS, D. M. (1980). Robust estimation of the variogram. I. *J. Int. Assoc. Math. Geol.* **12** 115–125. MR0595404 https://doi.org/10.1007/BF01035243

DALTON, A. (2016). Dark Sky's hyperlocal weather app is now available on the web. Available at https://www.engadget.com/2016/09/20/dark-sky-hyperlocal-weather-app-desktop-web/. Accessed: 2019-01-30.

FLORIO, E. N., LELE, S. R., CHANG, Y. C., STERNER, R. and GLASS, G. E. (2004). Integrating AVHRR satellite data and NOAA ground observations to predict surface air temperature: A statistical approach. *Int. J. Remote Sens.* **25** 2979–2994.

FREI, C. (2014). Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances. *Int. J. Climatol.* **34** 1585–1605.

GANDIN, L. S. (1988). Complex quality control of meteorological observations. *Mon. Weather Rev.* **116** 1137–1156.

GELFAND, A. E., DIGGLE, P. J., FUENTES, M. and GUTTORP, P. (2010). *Handbook of Spatial Statistics*. CRC Press, Boca Raton, FL.

GENTON, M. G. (1998). Highly robust variogram estimation. *Math. Geol.* **30** 213–221. MR1610687 https://doi.org/10.1023/A:1021728614555

Key words and phrases. Veracity score, geostatistics, robust kriging, hyperlocal spatial prediction.

GHOSH, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *Ann. Math. Stat.* **42** 1957–1961. MR0297071 https://doi.org/10.1214/aoms/1177693063

GNEITING, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19** 1327–1349. MR3102554 https://doi.org/10.3150/12-BEJSP06

HALL, P. and PATIL, P. (1994). Properties of nonparametric estimators of autocovariance for stationary random fields. *Probab. Theory Related Fields* **99** 399–424. MR1283119 https://doi.org/10.1007/BF01199899

HARRIS, P., BRUNSDON, C., CHARLTON, M., JUGGINS, S. and CLARKE, A. (2014). Multivariate spatial outlier detection using robust geographically weighted methods. *Math. Geosci.* **46** 1–31. MR3158063 https://doi.org/10.1007/s11004-013-9491-0

HASKARD, K. A. (2007). An anisotropic Matérn spatial covariance model: REML estimation and properties, Ph.D. thesis, Univ. Adelaide.

HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. Wiley, Hoboken, NJ. MR2488795 https://doi.org/10.1002/9780470434697

KOLLER, M. and STAHEL, W. A. (2011). Sharpening Wald-type inference in robust regression for small samples. *Comput. Statist. Data Anal.* **55** 2504–2515. MR2787008 https://doi.org/10.1016/j.csda.2011.02.014

KÜNSCH, H. R., PAPRITZ, A., SCHWIERZ, C. and STAHEL, A. W. (2011). Robust estimation of the external drift and the variogram of spatial data. In *ISI 58th World Statistics Congress of the International Statistical Institute, Dublin, Ireland* 21–26.

LAHIRI, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York. MR2001447 https://doi.org/10.1007/978-1-4757-3803-2

LAHIRI, S. N., LEE, Y. and CRESSIE, N. (2002). On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. *J. Statist. Plann. Inference* **103** 65–85. MR1896984 https://doi.org/10.1016/S0378-3758(01)00198-7

LARK, R. M. (2000). A comparison of some robust estimators of the variogram for use in soil survey. *Eur. J. Soil Sci.* **51** 137–157.

LORENC, A. C. (1986). Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.* **112** 1177–1194.

LUSSANA, C., UBOLDI, F. and SALVATI, M. R. (2010). A spatial consistency test for surface observations from mesoscale meteorological networks. *Q. J. R. Meteorol. Soc.* **136** 1075–1088.

MATHERON, G. (1962). *Traité de géostatistique appliquée, Tome I. Memoires du BRGM* (*Paris*) **14**. Technip, Paris.

MOYNIHAN, T. (2015). Clever app turns everyone into a roving weather reporter. Available at https://www.wired.com/2015/10/clever-app-turns-everyone-roving-weather-reporter/. Accessed: 2019-01-30.

PAPRITZ, A. (2018a). Tutorial and manual for geostatistical analyses with the R package georob. Available at https://cran.r-project.org/web/packages/georob/vignettes/georob_vignette.pdf. Accessed: 2019-02-12.

PAPRITZ, A. (2018b). georob: Robust geostatistical analysis of spatial data. R package version 0.3-7.

SEN, P. K. (1968). Asymptotic normality of sample quantiles for m-dependent processes. *Ann. Math. Stat.* **39** 1724–1730. MR0232522 https://doi.org/10.1214/aoms/1177698155

SOSKO, S. and DALYOT, S. (2017). Crowdsourcing user-generated mobile sensor weather data for densifying static geosensor networks. *ISPRS Int.l J. Geo-Inf.* **6** 61.

SUN, S. and LAHIRI, S. N. (2006). Bootstrapping the sample quantile of a weakly dependent sequence. *Sankhyā* **68** 130–166. MR2301568

THORNTON, P. E., RUNNING, S. W. and WHITE, M. A. (1997). Generating surfaces of daily meteorological variables over large regions of complex terrain. *J. Hydrol.* **1905** 214–251.

TODOROV, V. and FILZMOSER, P. (2009). An object-oriented framework for robust multivariate analysis. *J. Stat. Softw.* **32** 1–47.

VANCUTSEM, C., CECCATO, P., DINKU, T. and CONNOR, S. J. (2010). Evaluation of MODIS land surface temperature data to estimate air temperature in different ecosystems over Africa. *Remote Sens. Environ.* **114** 449–465.

WILLET, J. B. and SINGER, J. D. (1988). Another cautionary note about R^2: Its use in weighted least-square regression analysis. *Amer. Statist.* **42** 236–238.

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 143–159
https://doi.org/10.1214/19-AOAS1297

SURFACE TEMPERATURE MONITORING IN LIVER PROCUREMENT VIA FUNCTIONAL VARIANCE CHANGE-POINT ANALYSIS

BY ZHENGUO GAO[1], PANG DU[2], RAN JIN[3] AND JOHN L. ROBERTSON[4]

[1]*School of Mathematical Sciences, Shanghai Jiao Tong University, gaozhenguo3@126.com*
[2]*Department of Statistics, Virginia Tech, pangdu@vt.edu*
[3]*Grado Department of Industrial and Systems Engineering, Virginia Tech, jran5@vt.edu*
[4]*School of Biomedical Engineering, Virginia Tech, drbob@vt.edu*

Liver procurement experiments with surface-temperature monitoring motivated Gao et al. (*J. Amer. Statist. Assoc.* **114** (2019) 773–781) to develop a variance change-point detection method under a smoothly-changing mean trend. However, the spotwise change points yielded from their method do not offer immediate information to surgeons since an organ is often transplanted as a whole or in part. We develop a new practical method that can analyze a defined portion of the organ surface at a time. It also provides a novel addition to the developing field of functional data monitoring. Furthermore, numerical challenge emerges for simultaneously modeling the variance functions of 2D locations and the mean function of location and time. The respective sample sizes in the scales of 10,000 and 1,000,000 for modeling these functions make standard spline estimation too costly to be useful. We introduce a multistage subsampling strategy with steps educated by quickly-computable preliminary statistical measures. Extensive simulations show that the new method can efficiently reduce the computational cost and provide reasonable parameter estimates. Application of the new method to our liver surface temperature monitoring data shows its effectiveness in providing accurate status change information for a selected portion of the organ in the experiment.

REFERENCES

BERKES, I., GABRYS, R., HORVÁTH, L. and KOKOSZKA, P. (2009). Detecting changes in the mean of functional observations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 927–946. MR2750251 https://doi.org/10.1111/j.1467-9868.2009.00713.x

BHONSLE, S., BONAKDAR, M., NEAL II, R. E., AARDEMA, C., ROBERTSON, J. L., HOWARTH, J., KAVNOUDIAS, H., THOMSON, K. R., GOLDBERG, S. N. et al. (2016). Characterization of irreversible electroporation ablation with a validated perfused organ model. *J. Vasc. Interv. Radiol.* **27** 1913–1922.

DEMMY, T. L., BIDDLE, J. S., BENNETT, L. E., WALLS, J. T., SCHMALTZ, R. A. and CURTIS, J. J. (1997). Organ preservation solutions in heart transplantation-patterns of usage and related survival. *Transplant.* **63** 262–269.

DRINEAS, P., MAHONEY, M. W., MUTHUKRISHNAN, S. and SARLÓS, T. (2011). Faster least squares approximation. *Numer. Math.* **117** 219–249. MR2754850 https://doi.org/10.1007/s00211-010-0331-6

DUCHON, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables* (*Proc. Conf., Math. Res. Inst., Oberwolfach*, 1976) *Lecture Notes in Math.* **571** 85–100. Springer, Berlin. MR0493110

FAN, J., ZHANG, C. and ZHANG, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29** 153–193. MR1833962 https://doi.org/10.1214/aos/996986505

FEBRERO, M., GALEANO, P. and GONZÁLEZ-MANTEIGA, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NO_x levels. *Environmetrics* **19** 331–345. MR2440036 https://doi.org/10.1002/env.878

FITHIAN, W. and HASTIE, T. (2014). Local case-control sampling: Efficient subsampling in imbalanced data sets. *Ann. Statist.* **42** 1693–1724. MR3257627 https://doi.org/10.1214/14-AOS1220

Key words and phrases. Functional variance change point, thin-plate splines; subsampling, functional data, liver procurement.

GAO, Z., SHANG, Z., DU, P. and ROBERTSON, J. L. (2019). Variance change point detection under a smoothly-changing mean trend with application to liver procurement. *J. Amer. Statist. Assoc.* **114** 773–781. MR3963179 https://doi.org/10.1080/01621459.2018.1442341

GAO, Z., DU, P., JIN, R. and ROBERTSON, J. L. (2020). Supplement to "Surface temperature monitoring in liver procurement via functional variance change-point analysis." https://doi.org/10.1214/19-AOAS1297SUPP.

GU, C. (2013). *Smoothing Spline ANOVA Models*, 2nd ed. *Springer Series in Statistics* **297**. Springer, New York. MR3025869 https://doi.org/10.1007/978-1-4614-5369-7

JIN, R., CHANG, C. J. and SHI, J. (2011). Sequential measurement strategy for wafer geometric profile estimation. *IIE Trans.* **44** 1–12.

KARPELOWSKY, J. S. J. (2014). Near-infrared spectroscopy for monitoring renal transplant perfusion. *Pediatric Nephrology* **29** 2241–2242.

KEEFFE, E. B. (2001). Liver transplantation: Current status and novel approaches to liver replacement. *Gastroenterol.* **120** 749–762.

KIM, Y.-J. and GU, C. (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 337–356. MR2062380 https://doi.org/10.1046/j.1369-7412.2003.05316.x

KOCHAN, K., MASLAK, E., CHLOPICKI, S. and BARANSKA, M. (2015). FT-IR imaging for quantitative determination of liver fat content in non-alcoholic fatty liver. *Analyst* **140** 4997–5002.

LAN, Q., JIN, R. and ROBERTSON, J. L. (2015). Quantitative and qualitative evaluation for organ preservation in transplant. In *IIE Annual Conference*, *Proceedings* 2229–2236.

LAN, Q., SUN, H., ROBERTSON, J., DENG, X. and JIN, R. (2018). Non-invasive assessment of liver quality in transplantation based on thermal imaging analysis. *Comput. Methods Programs Biomed.* **164** 31–47.

LIN, X. and CARROLL, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J. Amer. Statist. Assoc.* **96** 1045–1056. MR1947252 https://doi.org/10.1198/016214501753208708

MA, P., MAHONEY, M. W. and YU, B. (2015). A statistical perspective on algorithmic leveraging. *J. Mach. Learn. Res.* **16** 861–911. MR3361306

MA, P. and SUN, X. (2015). Leveraging for big data regression. *Wiley Interdiscip. Rev.: Comput. Stat.* **7** 70–76. MR3348722 https://doi.org/10.1002/wics.1324

O'BRIEN, T. J., ROGHANIZAD, A. R., JONES, P. A., AARDEMA, C. H., ROBERTSON, J. L. and DILLER, T. E. (2017). The development of a thin-filmed noninvasive tissue perfusion sensor to quantify capillary pressure occlusion of explanted organs. *IEEE Trans. Biomed. Eng.* **64** 1631–1637.

QIU, P., ZOU, C. and WANG, Z. (2010). Nonparametric profile monitoring by mixed effects modeling. *Technometrics* **52** 265–277. MR2723706 https://doi.org/10.1198/TECH.2010.08188

QUAN, A., LEUNG, S. W., LAO, T. T. and MAN, R. Y. (2003). 5-hydroxytryptamine and thromboxane A2 as physiologic mediators of human umbilical artery closure. *J. Soc. Gynecol. Investig.* **10** 490–495.

ROKHLIN, V. and TYGERT, M. (2008). A fast randomized algorithm for overdetermined linear least-squares regression. *Proc. Natl. Acad. Sci. USA* **105** 13212–13217. MR2443725 https://doi.org/10.1073/pnas.0804869105

ROTHUIZEN, J. and TWEDT, D. C. (2009). Liver biopsy techniques. *Vet. Clin. North Am., Small Anim. Pract.* **39** 469–480.

SHANG, Z. and CHENG, G. (2017). Computational limits of a distributed algorithm for smoothing spline. *J. Mach. Learn. Res.* **18** Paper No. 108, 37. MR3725447

VARGAFTIG, B. B. and HAI, N. D. (1972). Selective inhibition by mepacrine of the release of "rabbit aorta contracting substance" evoked by the administration of bradykinin. *J. Pharm. Pharmacol.* **24** 159–161.

VAZQUEZ-MARTUL, E. and PAPADIMITRIOU, J. C. (2004). Importance of biopsy evaluation and the role of the pathologist in solid organ transplant programs. *Transplant. Proc.* **36** 725–728.

VIDAL, E., AMIGONI, A., BRUGNOLARO, V., GHIRARDO, G., GAMBA, P., PETTENAZZO, A., ZANON, G. F., COSMA, C., PLEBANI, M. et al. (2014). Nearinfrared spectroscopy as continuous real-time monitoring for kidney graft perfusion. *Pediatric Nephrology* **29** 909–914.

WAHBA, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442 https://doi.org/10.1137/1.9781611970128

WANG, H., YANG, M. and STUFKEN, J. (2019). Information-based optimal subdata selection for big data linear regression. *J. Amer. Statist. Assoc.* **114** 393–405. MR3941263 https://doi.org/10.1080/01621459.2017.1408468

WANG, H., ZHU, R. and MA, P. (2018). Optimal subsampling for large sample logistic regression. *J. Amer. Statist. Assoc.* **113** 829–844. MR3832230 https://doi.org/10.1080/01621459.2017.1292914

WOODALL, W. H., SPITZNER, D. J., MONTGOMERY, D. C. and GUPTA, S. (2004). Using control charts to monitor process and product profiles. *J. Qual. Technol.* **36** 309–320.

XU, D. and WANG, Y. (2018). Divide and recombine approaches for fitting smoothing spline models with large datasets. *J. Comput. Graph. Statist.* **27** 677–683. MR3863768 https://doi.org/10.1080/10618600.2017.1402775

YU, G., ZOU, C. and WANG, Z. (2012). Outlier detection in functional observations with applications to profile monitoring. *Technometrics* **54** 308–318. MR2967980 https://doi.org/10.1080/00401706.2012.694781

ZHANG, Y., DUCHI, J. and WAINWRIGHT, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.* **16** 3299–3340. MR3450540

ZOU, C., TSUNG, F. and WANG, Z. (2008). Monitoring profiles based on nonparametric regression methods. *Technometrics* **50** 512–526. MR2477862 https://doi.org/10.1198/004017008000000433

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 160–177
https://doi.org/10.1214/19-AOAS1304

ASSESSING WAGE STATUS TRANSITION AND STAGNATION USING QUANTILE TRANSITION REGRESSION

BY CHIH-YUAN HSU[1], YI-HAU CHEN[2], RUOH-RONG YU[3,*] AND TSUNG-WEI HUNG[3]

[1]*Department of Biostatistics, Vanderbilt University Medical Center, chih-yuan.hsu@vumc.org*
[2]*Institute of Statistical Science, Academia Sinica, yhchen@stat.sinica.edu.tw*
[3]*Center for Survey Research, Research Center for Humanities and Social Sciences, Academia Sinica,*
[*]*yurr@gate.sinica.edu.tw*

Workers in Taiwan overall have been suffering from long-lasting wage stagnation since the mid-1990s. In particular, there seems to be little mobility for the wages of Taiwanese workers to transit across wage quantile groups. It is of interest to see if certain groups of workers, such as female, lower educated and younger generation workers, suffer from the problem more seriously than the others. This work tries to apply a systematic statistical approach to study this issue, based on the longitudinal data from the Panel Study of Family Dynamics (PSFD) survey conducted in Taiwan since 1999. We propose the quantile transition regression model, generalizing recent methodology for quantile association, to assess the wage status transition with respect to the marginal wage quantiles over time as well as the effects of certain demographic and job factors on the wage status transition. Estimation of the model can be based on the composite likelihoods utilizing the binary, or ordinal-data information regarding the quantile transition, with the associated asymptotic theory established. A goodness-of-fit procedure for the proposed model is developed. The performances of the estimation and the goodness-of-fit procedures for the quantile transition model are illustrated through simulations. The application of the proposed methodology to the PSFD survey data suggests that female, private-sector workers with higher age and education below postgraduate level suffer from more severe wage status stagnation than the others.

REFERENCES

BROWN, B. M. and WANG, Y.-G. (2005). Standard errors and covariance matrices for smoothed rank estimators. *Biometrika* **92** 149–158. MR2158616 https://doi.org/10.1093/biomet/92.1.149

CAREY, V. C., ZEGER, S. L. and DIGGLE, P. J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80** 517–526.

CHEN, H.-J. and KUO, C.-W. (2014). The possible causes and solutions for Taiwan's wage stagnation. In *The R.O.C. Economy Yearbook* (Economic Daily News, ed.) 32–39. Economic Daily News, Taipei.

CHERNOZHUKOV, V. (2005). Extremal quantile regression. *Ann. Statist.* **33** 806–839. MR2163160 https://doi.org/10.1214/009053604000001165

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

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

FRUMENTO, P. and BOTTAI, M. (2016). Parametric modeling of quantile regression coefficient functions. *Biometrics* **72** 74–84. MR3500576 https://doi.org/10.1111/biom.12410

GALVAO, A. F. JR. (2011). Quantile regression for dynamic panel data with fixed effects. *J. Econometrics* **164** 142–157. MR2821799 https://doi.org/10.1016/j.jeconom.2011.02.016

HAEPP, T. and HSIN, P.-L. (2016). Is Taiwan's workforce underpaid? Evidence from marginal product of labor estimates at the company level. *J. Soc. Sci. Philos.* **28** 299–331.

Key words and phrases. Longitudinal data, panel study, quantile association, quantile regression, transition probability.

HEAGERTY, P. J. and ZEGER, S. L. (1996). Marginal regression models for clustered ordinal measurements. *J. Amer. Statist. Assoc.* **91** 1024–1036.

HSU, C.-Y., CHEN, Y.-H., YU, R.-R. and HUNG, T.-W. (2020). Supplement to "Assessing wage status transition and stagnation using quantile transition regression." https://doi.org/10.1214/19-AOAS1304SUPP.

HUANG, D.-S., LIU, B.-J. and YANG, T.-H. (2014). The phenomena of Taiwan's real wage stagnation: Global trend and Taiwanese characteristics. IEAS Working Paper No. 14-A012.

KENDALL, M. G. and STUART, A. (1973). *The Advanced Theory of Statistics. Vol.* 2: *Inference and Relationship*, 3rd ed. Hafner, New York. MR0474561

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

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

KUK, A. Y. C. (2007). A hybrid pairwise likelihood method. *Biometrika* **94** 939–952. MR2416800 https://doi.org/10.1093/biomet/asm051

LI, R., CHENG, Y. and FINE, J. P. (2014). Quantile association regression models. *J. Amer. Statist. Assoc.* **109** 230–242. MR3180559 https://doi.org/10.1080/01621459.2013.847375

LI, C.-T. and FANG, C.-T. (2015). *The Causes of the Decoupling of Taiwan's Labor Productivity and Wages. Presented at Conference of the Taiwanese Economic Associations*. National Taiwan Univ., Taipei.

LIN, C.-C., CHANG, J.-J. and LU, S.-S. (2017). Wage stagnation? Fact disclosure and cross-country comparison. *J. Soc. Sci. Philos.* **29** 87–125.

YANG, C.-C., CHEN, Y.-H. and CHANG, H.-Y. (2017). Joint regression analysis of marginal quantile and quantile association: Application to longitudinal body mass index in adolescents. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 1075–1090. MR3715598 https://doi.org/10.1111/rssc.12214

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 178–201
https://doi.org/10.1214/19-AOAS1302

TFISHER: A POWERFUL TRUNCATION AND WEIGHTING PROCEDURE FOR COMBINING *p*-VALUES

BY HONG ZHANG[1], TIEJUN TONG[2], JOHN LANDERS[3] AND ZHEYANG WU[4]

[1]*Merck Research Laboratories, hong.zhang8@merck.com*
[2]*Hong Kong Baptist University, tongt@hkbu.edu.hk*
[3]*University of Massachusetts Medical School, John.Landers@umassmed.edu*
[4]*Worcester Polytechnic Institute, zheyangwu@wpi.edu*

The *p*-value combination approach is an important statistical strategy for testing global hypotheses with broad applications in signal detection, meta-analysis, data integration, etc. In this paper we extend the classic Fisher's combination method to a unified family of statistics, called TFisher, which allows a general truncation-and-weighting scheme of input *p*-values. TFisher can significantly improve statistical power over the Fisher and related truncation-only methods for detecting both rare and dense "signals." To address wide applications, analytical calculations for TFisher's size and power are deduced under any two continuous distributions in the null and the alternative hypotheses. The corresponding omnibus test (oTFisher) and its size calculation are also provided for data-adaptive analysis. We study the asymptotic optimal parameters of truncation and weighting based on Bahadur efficiency (BE). A new asymptotic measure, called the asymptotic power efficiency (APE), is also proposed for better reflecting the statistics' performance in real data analysis. Interestingly, under the Gaussian mixture model in the signal detection problem, both BE and APE indicate that the soft-thresholding scheme is the best, the truncation and weighting parameters should be equal. By simulations of various signal patterns, we systematically compare the power of statistics within TFisher family as well as some rare-signal-optimal tests. We illustrate the use of TFisher in an exome-sequencing analysis for detecting novel genes of amyotrophic lateral sclerosis. Relevant computation has been implemented into an R package *TFisher* published on the Comprehensive R Archive Network to cater for applications.

REFERENCES

ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. L. and JOHNSTONE, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653. MR2281879 https://doi.org/10.1214/009053606000000074

ABU-DAYYEH, W. A., AL-MOMANI, M. A. and MUTTLAK, H. A. (2003). Exact Bahadur slope for combining independent tests for normal and logistic distributions. *Appl. Math. Comput.* **135** 345–360. MR1937258 https://doi.org/10.1016/S0096-3003(01)00336-8

ANDRÉS-BENITO, P., MORENO, J., ASO, E., POVEDANO, M. and FERRER, I. (2017). Amyotrophic lateral sclerosis, gene deregulation in the anterior horn of the spinal cord and frontal cortex area 8: Implications in frontotemporal lobar degeneration. *Aging* **9** 823–851. https://doi.org/10.18632/aging.101195

ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. MR2906877 https://doi.org/10.1214/11-AOS910

AYERS, K. L., MIRSHAHI, U. L., WARDEH, A. H., MURRAY, M. F., HAO, K., GLICKSBERG, B. S., LI, S., CAREY, D. J. and CHEN, R. (2016). A loss of function variant in CASP7 protects against Alzheimer's disease in homozygous APOE $\varepsilon 4$ allele carriers. *BMC Genomics* **17** 445.

AZZALINI, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Stat.* **12** 171–178. MR0808153

Key words and phrases. *P*-value combination, global hypothesis testing, signal detection, statistical power, optimal test, genetic association studies.

BAHADUR, R. R. (1960). Stochastic comparison of tests. *Ann. Math. Stat.* **31** 276–295. MR0116413 https://doi.org/10.1214/aoms/1177705894

BARNETT, I. J. and LIN, X. (2014). Analytical p-value calculation for the higher criticism test in finite-d problems. *Biometrika* **101** 964–970. MR3286929 https://doi.org/10.1093/biomet/asu033

BIERNACKA, J. M., JENKINS, G. D., WANG, L., MOYER, A. M. and FRIDLEY, B. L. (2012). Use of the gamma method for self-contained gene-set analysis of SNP data. *Eur. J. Hum. Genet.* **20** 565–571.

BONIFATI, V. (2006). Parkinson's disease: The LRRK2-G2019S mutation: Opening a novel era in Parkinson's disease genetics. *Eur. J. Hum. Genet.* **14** 1061–1062.

BRUCE, A. G. and GAO, H.-Y. (1996). Understanding WaveShrink: Variance and bias estimation. *Biometrika* **83** 727–745. MR1440040 https://doi.org/10.1093/biomet/83.4.727

CAI, T. T. and WU, Y. (2014). Optimal detection of sparse mixtures against a given null distribution. *IEEE Trans. Inform. Theory* **60** 2217–2232. MR3181520 https://doi.org/10.1109/TIT.2014.2304295

CARTER, B. J., ANKLESARIA, P., CHOI, S. and ENGELHARDT, J. F. (2009). Redox modifier genes and pathways in amyotrophic lateral sclerosis. *Antioxid. Redox Signal.* **11** 1569–1586.

CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference*, 2nd ed. Duxbury, Pacific Grove, CA.

CEVIKBAS, F., WANG, X., AKIYAMA, T., KEMPKES, C., SAVINKO, T., ANTAL, A., KUKOVA, G., BUHL, T., IKOMA, A. et al. (2014). A sensory neuron–expressed IL-31 receptor mediates T helper cell–dependent itch: Involvement of TRPV1 and TRPA1. *J. Allergy Clin. Immunol.* **133** 448–460.

CHAPMAN, D. L. and PAPAIOANNOU, V. E. (1998). Three neural tubes in mouse embryos with mutations in the T-box gene Tbx6. *Nature* **391** 695–697.

CHEN, C.-W. and YANG, H.-C. (2017). OPATs: Omnibus P-value association tests. *Brief. Bioinform.* **20** 1–14.

COX, L. E., FERRAIUOLO, L., GOODALL, E. F., HEATH, P. R., HIGGINBOTTOM, A., MORTIBOYS, H., HOLLINGER, H. C., HARTLEY, J. A., BROCKINGTON, A. et al. (2010). Mutations in CHMP2B in lower motor neuron predominant amyotrophic lateral sclerosis (ALS). *PLoS ONE* **5** e9872.

DAI, H., LEEDER, J. S. and CUI, Y. (2014). A modified generalized Fisher method for combining probabilities from dependent tests. *Front. Genet.* **5** 32. https://doi.org/10.3389/fgene.2014.00032

DANIELS, H. E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Stat.* **25** 631–650. MR0066602 https://doi.org/10.1214/aoms/1177728652

DASGUPTA, A. (2008). *Asymptotic Theory of Statistics and Probability*. *Springer Texts in Statistics*. Springer, New York. MR2664452

DE OLIVEIRA, G. P., MAXIMINO, J. R., MASCHIETTO, M., ZANOTELI, E., PUGA, R. D., LIMA, L., CARRARO, D. M. and CHADI, G. (2014). Early gene expression changes in skeletal muscle from SOD1G93A amyotrophic lateral sclerosis animal model. *Cell. Mol. Neurobiol.* **34** 451–462.

DONOHO, D. L. (1995). De-noising by soft-thresholding. *IEEE Trans. Inform. Theory* **41** 613–627. MR1331258 https://doi.org/10.1109/18.382009

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089 https://doi.org/10.1093/biomet/81.3.425

DUDBRIDGE, F. and KOELEMAN, B. P. C. (2003). Rank truncated product of P-values, with application to genomewide association scans. *Genet. Epidemiol.* **25** 360–366.

DUERR, R. H., TAYLOR, K. D., BRANT, S. R., RIOUX, J. D., SILVERBERG, M. S., DALY, M. J., STEINHART, A. H., ABRAHAM, C., REGUEIRO, M. et al. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Sci. Signal.* **314** 1461.

FANNING, S., XU, W., BEAUREPAIRE, C., SUHAN, J. P., NANTEL, A. and MITCHELL, A. P. (2012). Functional control of the *Candida albicans* cell wall by catalytic protein kinase A subunit Tpk1. *Mol. Microbiol.* **86** 284–302.

FISHER, R. A. (1932). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

GENZ, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comput. Graph. Statist.* **1** 141–149.

GOOD, I. J. (1955). Ont the wieghted combination of significance tests. *J. Roy. Statist. Soc. Ser. B* **17** 264–265. MR0076252

GUO, S., LI, Z.-Z., GONG, J., XIANG, M., ZHANG, P., ZHAO, G.-N., LI, M., ZHENG, A., ZHU, X. et al. (2015). Oncostatin M confers neuroprotection against ischemic stroke. *J. Neurosci.* **35** 12047–12062.

HOH, J., WILLE, A. and OTT, J. (2001). Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res.* **11** 2115–2119.

INGSTER, Y. I. (2002). Adaptive detection of a signal of growing dimension. II. *Math. Methods Statist.* **11** 37–68. MR1900973

INGSTER, Y. I., TSYBAKOV, A. B. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526. MR2747131 https://doi.org/10.1214/10-EJS589

KUO, C.-L. and ZAYKIN, D. V. (2011). Novel rank-based approaches for discovery and replication in genome-wide association studies. *Genetics* **189** 329–340.
LEE, S., EMOND, M. J., BAMSHAD, M. J., BARNES, K. C., RIEDER, M. J., NICKERSON, D. A., CHRISTIANI, D. C., WURFEL, M. M. and LIN, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91** 224–237.
LI, J. and TSENG, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann. Appl. Stat.* **5** 994–1019. MR2840184 https://doi.org/10.1214/10-AOAS393
LIN, X., LEE, S., WU, M. C., WANG, C., CHEN, H., LI, Z. and LIN, X. (2016). Test for rare variants by environment interactions in sequencing association studies. *Biometrics* **72** 156–164. MR3500584 https://doi.org/10.1111/biom.12368
LITTELL, R. C. and FOLKS, J. L. (1971). Asymptotic optimality of Fisher's method of combining independent tests. *J. Amer. Statist. Assoc.* **66** 802–806. MR0312634
LITTELL, R. C. and FOLKS, J. L. (1973). Asymptotic optimality of Fisher's method of combining independent tests. II. *J. Amer. Statist. Assoc.* **68** 193–194. MR0375577
LUGANNANI, R. and RICE, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Adv. in Appl. Probab.* **12** 475–490. MR0569438 https://doi.org/10.2307/1426607
MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6
MORAHAN, J. M., YU, B., TRENT, R. J. and PAMPHLETT, R. (2009). A genome-wide analysis of brain DNA methylation identifies new candidate genes for sporadic amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler.* **10** 418–429.
NADARAJAH, S. (2005). A generalized normal distribution. *J. Appl. Stat.* **32** 685–694. MR2119411 https://doi.org/10.1080/02664760500079464
NIKITIN, Y. (1995). *Asymptotic Efficiency of Nonparametric Tests*. Cambridge Univ. Press, Cambridge. MR1335235 https://doi.org/10.1017/CBO9780511530081
SCHAID, D. J., ROWLAND, C. M., TINES, D. E., JACOBSON, R. M. and POLAND, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* **70** 425–434.
SMITH, B. N., TICOZZI, N., FALLINI, C., GKAZI, A. S., TOPP, S., KENNA, K. P., SCOTTER, E. L., KOST, J., KEAGLE, P. et al. (2014). Exome-wide rare variant analysis identifies TUBA4A mutations associated with familial ALS. *Neuron* **84** 324–331.
SONG, C. and TSENG, G. C. (2014). Hypothesis setting and order statistic for robust genomic meta-analysis. *Ann. Appl. Stat.* **8** 777–800. MR3262534 https://doi.org/10.1214/13-AOAS683
STOUFFER, S. A., SUCHMAN, E. A., DEVINNEY, L. C., STAR, S. A. and WILLIAMS, R. M. (1949). *The American Soldier*: *Adjustment During Army Life* **I**. Princeton Univ. Press, Princeton, NJ.
SU, Y.-C., GAUDERMAN, W. J., BERHANE, K. and LEWINGER, J. P. (2016). Adaptive set-based methods for association testing. *Genet. Epidemiol.* **40** 113–122.
SZKLARCZYK, D., FRANCESCHINI, A., WYDER, S., FORSLUND, K., HELLER, D., HUERTA-CEPAS, J., SIMONOVIC, M., ROTH, A., SANTOS, A. et al. (2014). STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43** D447–D452.
PANDYA, S., MAO, L. L., ZHOU, E. W., BOWSER, R., ZHU, Z., ZHU, Y. and WANG, X. (2012). Neuroprotection for amyotrophic lateral sclerosis: Role of stem cells, growth factors, and gene therapy. *Cent. Nerv. Syst. Agents. Med. Chem.* **12** 15–27.
VARANASI, M. K. and AAZHANG, B. (1989). Parametric generalized Gaussian density estimation. *J. Acoust. Soc. Am.* **86** 1404–1415.
WHITLOCK, M. C. (2005). Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **18** 1368–1373.
WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89** 82–93.
WU, Z., SUN, Y., HE, S., CHO, J., ZHAO, H. and JIN, J. (2014). Detection boundary and higher criticism approach for rare and weak genetic effects. *Ann. Appl. Stat.* **8** 824–851. MR3262536 https://doi.org/10.1214/14-AOAS724
YU, K., LI, Q., BERGEN, A. W., PFEIFFER, R. M., ROSENBERG, P. S., CAPORASO, N., KRAFT, P. and CHATTERJEE, N. (2009). Pathway analysis by adaptive combination of P-values. *Genet. Epidemiol.* **33** 700–709.
ZAYKIN, D. V., ZHIVOTOVSKY, L. A., WESTFALL, P. H. and WEIR, B. S. (2002). Truncated product method for combining P-values. *Genet. Epidemiol.* **22** 170–185.
ZAYKIN, D. V., ZHIVOTOVSKY, L. A., CZIKA, W., SHAO, S. and WOLFINGER, R. D. (2007). Combining p-values in large-scale genomics experiments. *Pharm. Stat.* **6** 217–226.
ZHANG, J. and HUANG, E. J. (2006). Dynamic expression of neurotrophic factor receptors in postnatal spinal motoneurons and in mouse model of ALS. *J. Neurobiol.* **66** 882–895.

ZHANG, H., TONG, T., LANDERS, J. E. and WU, Z. (2020). Supplement to "TFisher: A truncation and weighting procedure for combining p-values." https://doi.org/10.1214/19-AOAS1302SUPP.

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 202–220
https://doi.org/10.1214/19-AOAS1301

MODIFYING THE CHI-SQUARE AND THE CMH TEST FOR POPULATION GENETIC INFERENCE: ADAPTING TO OVERDISPERSION

BY KERSTIN SPITZER[1,*], MARTA PELIZZOLA[1,†] AND ANDREAS FUTSCHIK[2]

[1]*Vienna Graduate School of Population Genetics, Vetmeduni Vienna,* [*]*kerstin.e.spitzer@gmail.com;* [†]*Marta.Pelizzola@vetmeduni.ac.at*

[2]*Department of Applied Statistics, Johannes Kepler University Linz, andreas.futschik@jku.at*

Evolve and resequence studies provide a popular approach to simulate evolution in the lab and explore its genetic basis. In this context, Pearson's chi-square test, Fisher's exact test as well as the Cochran–Mantel–Haenszel test are commonly used to infer genomic positions affected by selection from temporal changes in allele frequency. However, the null model associated with these tests does not match the null hypothesis of actual interest. Indeed, due to genetic drift and possibly other additional noise components such as pool sequencing, the null variance in the data can be substantially larger than accounted for by these common test statistics. This leads to p-values that are systematically too small and, therefore, a huge number of false positive results. Even, if the ranking rather than the actual p-values is of interest, a naive application of the mentioned tests will give misleading results, as the amount of overdispersion varies from locus to locus. We therefore propose adjusted statistics that take the overdispersion into account while keeping the formulas simple. This is particularly useful in genome-wide applications, where millions of SNPs can be handled with little computational effort. We then apply the adapted test statistics to real data from *Drosophila* and investigate how information from intermediate generations can be included when available. We also discuss further applications such as genome-wide association studies based on pool sequencing data and tests for local adaptation.

REFERENCES

AGRESTI, A. (2002). *Categorical Data Analysis*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, New York. MR1914507 https://doi.org/10.1002/0471249688

BARGHI, N., TOBLER, R., NOLTE, V. and SCHLÖTTERER, C. (2017). *Drosophila simulans*: A species with improved resolution in evolve and resequence studies. *G3*: *Genes*, *Genomes*, *Genetics* **7** 2337–2343. https://doi.org/10.1534/g3.117.043349

BASTIDE, H., BETANCOURT, A., NOLTE, V., TOBLER, R., STÖBE, P., FUTSCHIK, A. and SCHLÖTTERER, C. (2013). A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. *PLoS Genet*. **9** e1003534. https://doi.org/10.1371/journal.pgen.1003534.

BEAUMONT, M. A. and BALDING, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol*. *Ecol*. **13** 969–980. https://doi.org/10.1111/j.1365-294X.2004.02125.x.

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

BOLLBACK, J. P., YORK, T. L. and NIELSEN, R. (2008). Estimation of 2Nes from temporal allele frequency data. *Genetics* **179** 497–502. https://doi.org/10.1534/genetics.107.085019

BURKE, M. K., DUNHAM, J. P., SHAHRESTANI, P., THORNTON, K. R., ROSE, M. R. and LONG, A. D. (2010). Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* **467** 587–590. https://doi.org/10.1038/nature09352.

ENDLER, L., BETANCOURT, A. J., NOLTE, V. and SCHLÖTTERER, C. (2016). Reconciling differences in pool-gwas between populations: A case study of female abdominal pigmentation in *Drosophila melanogaster*. *Genetics* **202** 843–855. https://doi.org/10.1534/genetics.115.183376.

EWENS, W. J. (2004). *Mathematical Population Genetics*. *I*: *Theoretical Introduction*, 2nd ed. *Interdisciplinary Applied Mathematics* **27**. Springer, New York. MR2026891 https://doi.org/10.1007/978-0-387-21822-9

Key words and phrases. Chi-square test, CMH test, overdispersion, experimental evolution, evolve and resequence, genetic drift, pool sequencing.

FALCONER, D. S. (1960). *Introduction to Quantitative Genetics*. The Ronald Press Company, New York.
FEDER, A. F., KRYAZHIMSKIY, S. and PLOTKIN, J. B. (2014). Identifying signatures of selection in genetic time series. *Genetics* **196** 509–522. https://doi.org/10.1534/genetics.113.158220.
FOLL, M. and GAGGIOTTI, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* **180** 977–993. https://doi.org/10.1534/genetics.108.092221.
FOLL, M., SHIM, H. and JENSEN, J. D. (2015). WFABC: A Wright–Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Mol. Ecol.* **15** 87–98. https://doi.org/10.1111/1755-0998.12280.
GRIFFIN, P. C., HANGARTNER, S. B., FOURNIER-LEVEL, A. and HOFFMANN, A. A. (2017). Genomic trajectories to desiccation resistance: Convergence and divergence among replicate selected *Drosophila* lines. *Genetics* **205** 871–890. https://doi.org/10.1534/genetics.116.187104
ILLINGWORTH, C. J. R., PARTS, L., SCHIFFELS, S., LITI, G. and MUSTONEN, V. (2012). Quantifying selection acting on a complex trait using allele frequency time series data. *Mol. Biol. Evol.* **29** 1187–1197. https://doi.org/10.1093/molbev/msr289.
IRANMEHR, A., AKBARI, A., SCHLÖTTERER, C. and BAFNA, V. (2017). CLEAR: Composition of Likelihoods for Evolve and Resequence Experiments. *Genetics* 1011–1023. https://doi.org/10.1101/080085.
JÓNÁS, Á., TAUS, T., KOSIOL, C., SCHLÖTTERER, C. and FUTSCHIK, A. (2016). Estimating the effective population size from temporal allele frequency changes in experimental evolution. *Genetics* **204** 723–735. https://doi.org/10.1534/genetics.116.191197
KOFLER, R., PANDEY, R. V. and SCHLÖTTERER, C. (2011). PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27** 3435–3436. https://doi.org/10.1093/bioinformatics/btr589
KOFLER, R. and SCHLÖTTERER, C. (2014). A guide for the design of evolve and resequencing studies. *Mol. Biol. Evol.* **31** 474–483. https://doi.org/10.1093/molbev/mst221.
LEVY, S. F., BLUNDELL, J. R., VENKATARAM, S., PETROV, D. A., FISHER, D. S. and SHERLOCK, G. (2015). Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519** 181–186. https://doi.org/10.1038/nature14279.
LOHR, S. L. (2010). *Sampling*: *Design and Analysis*, 2nd ed. Brooks/Cole Cengage Learning, Boston, MA. MR3057878
MALASPINAS, A. S., MALASPINAS, O., EVANS, S. N. and SLATKIN, M. (2012). Estimating allele age and selection coefficient from time-serial data. *Genetics* **192** 599–607. https://doi.org/10.1534/genetics.112.140939.
MATHIESON, I. and MCVEAN, G. (2013). Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics* **193** 973–984. https://doi.org/10.1534/genetics.112.147611.
MCDONALD, J. H. (2014). *Handbook of Biological Statistics*, 3rd ed. Sparky House Publishing, Baltimore, MD.
NOUHAUD, P., TOBLER, R., NOLTE, V. and SCHLÖTTERER, C. (2016). Ancestral population reconstitution from isofemale lines as a tool for experimental evolution. *Ecol. Evol.* **6** 7169–7175. https://doi.org/10.1002/ece3.2402.
OROZCO-TERWENGEL, P., KAPUN, M., NOLTE, V., KOFLER, R., FLATT, T. and SCHLÖTTERER, C. (2012). Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Mol. Ecol.* **21** 4931–4941. https://doi.org/10.1111/j.1365-294X.2012.05673.x.
R-CORE-TEAM (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.
REMOLINA, S. C., CHANG, P. L., LEIPS, J., NUZHDIN, S. V. and HUGHES, K. A. (2012). Genomic basis of aging and life-history evolution in *Drosophila melanogaster*. *Evolution* **66** 3390–3403. https://doi.org/10.1111/j.1558-5646.2012.01710.x.
SCHRAIBER, J. G., EVANS, S. N. and SLATKIN, M. (2016). Bayesian inference of natural selection from allele frequency time series. *Genetics* **203** 493–511. https://doi.org/10.1534/genetics.116.187278.
SPITZER, K., PELIZZOLA, M. and FUTSCHIK, A. (2020). Supplement to "Modifying the Chi-square and the CMH test for population genetic inference: Adapting to overdispersion." https://doi.org/10.1214/19-AOAS1301SUPPA, https://doi.org/10.1214/19-AOAS1301SUPPB, https://doi.org/10.1214/19-AOAS1301SUPPC.
STEINRÜCKEN, M., BHASKAR, A. and SONG, Y. S. (2014). A novel spectral method for inferring general diploid selection from time series genetic data. *Ann. Appl. Stat.* **8** 2203–2222. MR3292494 https://doi.org/10.1214/14-AOAS764
TAUS, T., FUTSCHIK, A. and SCHLÖTTERER, C. (2017). Quantifying selection with pool-seq time series data. *Mol. Biol. Evol.* **34** 3023–3034. https://doi.org/10.1093/molbev/msx225.
TERHORST, J., SCHLÖTTERER, C. and SONG, Y. S. (2015). Multi-locus analysis of genomic time series data from experimental evolution. *PLoS Genet.* **11** e1005069. https://doi.org/10.1371/journal.pgen.1005069.

TOBLER, R., HERMISSON, J. and SCHLÖTTERER, C. (2015). Parallel trait adaptation across opposing thermal environments in experimental *Drosophila melanogaster* populations. *Evolution* **69** 1745–1759. https://doi.org/10.1111/evo.12705.

TOBLER, R., FRANSSEN, S. U., KOFLER, R., OROZCO-TERWENGEL, P., NOLTE, V., HERMISSON, J. and SCHLÖTTERER, C. (2014). Massive habitat-specific genomic response in *D. melanogaster* populations during experimental evolution in hot and cold environments. *Mol. Biol. Evol.* **31** 364–375. https://doi.org/10.1093/molbev/mst205.

TOPA, H., JÓNÁS, Á., KOFLER, R., KOSIOL, C. and HONKELA, A. (2015). Gaussian process test for high-throughput sequencing time series: Application to experimental evolution. *Bioinformatics* **31** 1762–1770.

TURNER, T. L. and MILLER, P. M. (2012). Investigating natural variation in *Drosophila* courtship song by the evolve and resequence approach. *Genetics* **191** 633–642. https://doi.org/10.1534/genetics.112.139337.

TURNER, T. L., STEWART, A. D., FIELDS, A. T., RICE, W. R. and TARONE, A. M. (2011). Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. *PLoS Genet.* **7** e1001336. https://doi.org/10.1371/journal.pgen.1001336.

VLACHOS, C., BURNY, C., PELIZZOLA, M., BORGES, R., FUTSCHIK, A., KOFLER, R. and SCHLÖTTERER, C. (2019). *Genome Biol.* **20** 169. https://doi.org/10.1186/s13059-019-1770-8.

WAPLES, R. S. (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* **121** 379–391.

WILSON, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proc. Natl. Acad. Sci. USA* **116** 1195–1200. MR3904688 https://doi.org/10.1073/pnas.1814092116

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 221–240
https://doi.org/10.1214/19-AOAS1296

A HIERARCHICAL BAYESIAN MODEL FOR PREDICTING ECOLOGICAL INTERACTIONS USING SCALED EVOLUTIONARY RELATIONSHIPS

BY MOHAMAD ELMASRI[1,*], MAXWELL J. FARRELL[2], T. JONATHAN DAVIES[3] AND DAVID A. STEPHENS[1,**]

[1]*Department of Mathematics and Statistics, McGill University,* [*]*mohamad.elmsari@mail.mcgill.ca;* [**]*david.stephens@mcgill.ca*

[2]*Department of Biology, McGill University, maxwell.farrell@mail.mcgill.ca*

[3]*Departments of Botany, and Forest and Conservation Sciences, University of British Columbia, j.davies@ubc.ca*

Identifying undocumented or potential future interactions among species is a challenge facing modern ecologists. Recent link prediction methods rely on trait data; however, large species interaction databases are typically sparse and covariates are limited to only a fraction of species. On the other hand, evolutionary relationships, encoded as phylogenetic trees, can act as proxies for underlying traits and historical patterns of parasite sharing among hosts. We show that, using a network-based conditional model, phylogenetic information provides strong predictive power in a recently published global database of host-parasite interactions. By scaling the phylogeny using an evolutionary model, our method allows for biological interpretation often missing from latent variable models. To further improve on the phylogeny-only model, we combine a hierarchical Bayesian latent score framework for bipartite graphs that accounts for the number of interactions per species with host dependence informed by phylogeny. Combining the two information sources yields significant improvement in predictive accuracy over each of the submodels alone. As many interaction networks are constructed from presence-only data, we extend the model by integrating a correction mechanism for missing interactions which proves valuable in reducing uncertainty in unobserved interactions.

REFERENCES

AGUIRRE, A. A., KEEFE, T. J., REIF, J. S., KASHINSKY, L., YOCHEM, P. K., SALIKI, J. T., STOTT, J. L., GOLDSTEIN, T., DUBEY, J. P. et al. (2007). Infectious disease monitoring of the endangered Hawaiian monk seal. *J. Wildl. Dis.* **43** 229–241.

ALBERT, R. and BARABÁSI, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys.* **74** 47–97. MR1895096 https://doi.org/10.1103/RevModPhys.74.47

BARTOMEUS, I. (2013). Understanding linkage rules in plant-pollinator networks by using hierarchical models that incorporate pollinator detectability and plant traits. *PLoS ONE* **8** e69200.

BASTAZINI, V. A. G., FERREIRA, P. M. A., AZAMBUJA, B. O., CASAS, G., DEBASTIANI, V. J., GUIMARÃES, P. R. and PILLAR, V. D. (2017). Untangling the tangled bank: A novel method for partitioning the effects of phylogenies and traits on ecological networks. *Evol. Biol.* **44** 312–324.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

BESAG, J. (1986). On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B* **48** 259–302. MR0876840

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

BRAGA, M. P., RAZZOLINI, E. and BOEGER, W. A. (2015). Drivers of parasite sharing among Neotropical freshwater fishes. *J. Anim. Ecol.* **84** 487–497.

BRIX, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. in Appl. Probab.* **31** 929–953. MR1747450 https://doi.org/10.1239/aap/1029955251

CARON, F. and FOX, E. B. (2017). Sparse graphs using exchangeable random measures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1295–1366. MR3731666 https://doi.org/10.1111/rssb.12233

Key words and phrases. Ecological networks, composite likelihood, iterated conditional modes, presence-only networks, link prediction.

CHIU, G. S. and WESTVELD, A. H. (2011). A unifying approach for food webs, phylogeny, social networks, and statistics. *Proc. Natl. Acad. Sci. USA* **108** 15881–15886.

CHUNG, F. and LU, L. (2006). *Complex Graphs and Networks*. *CBMS Regional Conference Series in Mathematics* **107**. Amer. Math. Soc., Providence, RI. MR2248695 https://doi.org/10.1090/cbms/107

CLEAVELAND, S., LAURENSON, M. K. and TAYLOR, L. H. (2001). Diseases of humans and their domestic mammals: Pathogen characteristics, host range and the risk of emergence. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **356** 991–999.

DALLAS, T., PARK, A. W. and DRAKE, J. M. (2017). Predicting cryptic links in host-parasite networks. *PLoS Comput. Biol.* **13** 1–15.

DAVIES, T. J. and PEDERSEN, A. B. (2008). Phylogeny and geography predict pathogen community similarity in wild primates and humans. *Proc. Biol. Sci.* **275** 1695–1701. https://doi.org/10.1098/rspb.2008.0284

DEMŠAR, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7** 1–30. MR2274360

EHM, W., GNEITING, T., JORDAN, A. and KRÜGER, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 505–562. MR3506792 https://doi.org/10.1111/rssb.12154

ELMASRI, M., FARRELL, M., DAVIES, T. J. and STEPHENS, D. A. (2020). Supplement to "A hierarchical Bayesian model for predicting ecological interactions using scaled evolutionary relationships." https://doi.org/10.1214/19-AOAS1296SUPP.

FARRELL, M. J., BERRANG-FORD, L. and DAVIES, T. J. (2013). The study of parasite sharing for surveillance of zoonotic diseases. *Environ. Res. Lett.* **8** 015036.

FARRELL, M. J., STEPHENS, P. R., BERRANG-FORD, L., GITTLEMAN, J. L. and DAVIES, T. J. (2015). The path to host extinction can lead to loss of generalist parasites. *J. Anim. Ecol.* **84** 978–984.

FRITZ, S. A., BININDA-EMONDS, O. R. P. and PURVIS, A. (2009). Geographical variation in predictors of mammalian extinction risk: Big is bad, but only in the tropics. *Ecol. Lett.* **12** 538–549.

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

GILBERT, G. S. and WEBB, C. O. (2007). Phylogenetic signal in plant pathogen-host range. *Proc. Natl. Acad. Sci. USA* **104** 4979–4983.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548 https://doi.org/10.1198/016214506000001437

GÓMEZ, J. M., VERDÚ, M. and PERFECTTI, F. (2010). Ecological interactions are evolutionarily conserved across the entire tree of life. *Nature* **465** 918–21.

GRAVEL, D., POISOT, T., ALBOUY, C., VELEZ, L. and MOUILLOT, D. (2013). Inferring food web structure from predator-prey body size relationships. *Methods Ecol. Evol.* **4** 1083–1090.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504 https://doi.org/10.2307/3318737

HARMON, L. J., LOSOS, J. B., JONATHAN DAVIES, T., GILLESPIE, R. G., GITTLEMAN, J. L., BRYAN JENNINGS, W., KOZAK, K. H., MCPEEK, M. A., MORENO-ROARK, F. et al. (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution* **64** 2385–2396.

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

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

HUANG, S., DRAKE, J. M., GITTLEMAN, J. L. and ALTIZER, S. (2015). Parasite diversity declines with host evolutionary distinctiveness: A global analysis of carnivores. *Evolution* **69** 621–630.

JIANG, X., GOLD, D. and KOLACZYK, E. D. (2011). Network-based auto-probit modeling for protein function prediction. *Biometrics* **67** 958–966. MR2829270 https://doi.org/10.1111/j.1541-0420.2010.01519.x

JORDANO, P. (2016). Sampling networks of ecological interactions. *Funct. Ecol.* **30** 1883–1893.

KRIVITSKY, P. N. and HANDCOCK, M. S. (2008). Fitting position latent cluster models for social networks with latentnet. *J. Stat. Softw.* **24**. https://doi.org/10.18637/jss.v024.i05

KRIVITSKY, P. N. and HANDCOCK, M. S. (2017). latentnet: Latent position and cluster models for statistical networks. The Statnet Project. R package version 2.8.0. Available at http://www.statnet.org.

LA SALLE, J., WILLIAMS, K. J. and MORITZ, C. (2016). Biodiversity analysis in the digital era. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **371** 20150337.

LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 715–740. MR2370077 https://doi.org/10.1111/j.1467-9868.2007.00609.x

LUIS, A. D., O'SHEA, T. J., HAYMAN, D. T. S., WOOD, J. L. N., CUNNINGHAM, A. A., GILBERT, A. T., MILLS, J. N. and WEBB, C. T. (2015). Network analysis of host-virus communities in bats and rodents reveals determinants of cross-species transmission. *Ecol. Lett.* **18** 1153–1162.

MORALES-CASTILLA, I., MATIAS, M. G., GRAVEL, D. and ARAÚJO, M. B. (2015). Inferring biotic interactions from proxies. *Trends Ecol. Evol.* **30** 347–356.

OLIVAL, K. J., HOSSEINI, P. R., ZAMBRANA-TORRELIO, C., ROSS, N., BOGICH, T. L. and DASZAK, P. (2017). Host and viral traits predict zoonotic spillover from mammals. *Nature* **546** 646–650. https://doi.org/10.1038/nature22975

OVASKAINEN, O., ABREGO, N., HALME, P. and DUNSON, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods Ecol. Evol.* **7** 549–555.

OVASKAINEN, O., TIKHONOV, G., NORBERG, A., GUILLAUME BLANCHET, F., DUAN, L., DUNSON, D., ROSLIN, T. and ABREGO, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* **20** 561–576.

PAGEL, M. (1999). Inferring the historical patterns of biological evolution. *Nature* **401** 877–884.

PARK, A., FARRELL, M., SCHMIDT, J., HUANG, S., DALLAS, T., PAPPALARDO, P., DRAKE, J., STEPHENS, P., POULIN, R. et al. (2018). Characterizing the phylogenetic specialism–generalism spectrum of mammal parasites. *Proc. R. Soc. Lond., B Biol. Sci.* **285** 20172613.

PARRISH, C. R., HOLMES, E. C., MORENS, D. M., PARK, E.-C., BURKE, D. S., CALISHER, C. H., LAUGHLIN, C. A., SAIF, L. J. and DASZAK, P. (2008). Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol. Mol. Biol. Rev.* **72** 457–70.

PEARSE, I. S. and ALTERMATT, F. (2013). Predicting novel trophic interactions in a non-native world. *Ecol. Lett.* **16** 1088–1094.

PEDERSEN, A. B., ALTIZER, S., POSS, M., CUNNINGHAM, A. A. and NUNN, C. L. (2005). Patterns of host specificity and transmission among parasites of wild primates. *Int. J. Parasitol.* **35** 647–657.

PEDERSEN, A. B., JONES, K. E., NUNN, C. L. and ALTIZER, S. (2007). Infectious diseases and extinction risk in wild mammals. *Conserv. Biol.* **21** 1269–1279.

PETCHEY, O. L., BECKERMAN, A. P., RIEDE, J. O. and WARREN, P. H. (2008). Size, foraging, and food web structure. *Proc. Natl. Acad. Sci. USA* **105** 4191–4196.

POELEN, J. H., SIMONS, J. D. and MUNGALL, C. J. (2014). Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecol. Inform.* **24** 148–159.

ROBERT, C. P. and CASELLA, G. (2013). *Monte Carlo Statistical Methods*. *Springer Texts in Statistics*. Springer, New York. MR1707311 https://doi.org/10.1007/978-1-4757-3071-5

STEPHENS, P. R., PAPPALARDO, P., HUANG, S., BYERS, J. E., FARRELL, M. J., GEHMAN, A., GHAI, R. R., HAAS, S. E., HAN, B. et al. (2017). Global mammal parasite database version 2.0. *Ecology* **98** 1476–1476.

STOCK, M., POISOT, T., WAEGEMAN, W. and DE BAETS, B. (2017). Linear filtering reveals false negatives in species interaction data. *Sci. Rep.* **7** 1–8.

STREICKER, D. G., TURMELLE, A. S., VONHOF, M. J., KUZMIN, I. V., MCCRACKEN, G. F. and RUPPRECHT, C. E. (2010). Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. *Science* **329** 676–679. https://doi.org/10.1126/science.1188836

SWENDSEN, R. H. and WANG, J.-S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58** 86–88.

TEH, Y. W. and GORUR, D. (2009). Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems* **22** 1838–1846.

WARDEH, M., RISLEY, C., MCINTYRE, M. K., SETZKORN, C. and BAYLIS, M. (2015). Database of host-pathogen and related species interactions, and their global distribution. *Sci. Data* **2** 150049. https://doi.org/10.1038/sdata.2015.49

WEBB, C. O., ACKERLY, D. D., MCPEEK, M. A. and DONOGHUE, M. J. (2002). Phylogenies and community ecology. *Ann. Rev. Ecolog. Syst.* **33** 475–505.

WEIR, I. S. and PETTITT, A. N. (2000). Binary probability maps using a hidden conditional autoregressive Gaussian process with an application to Finnish common toad data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **49** 473–484. MR1824553 https://doi.org/10.1111/1467-9876.00206

WIENS, J. J., ACKERLY, D. D., ALLEN, A. P., ANACKER, B. L., BUCKLEY, L. B., CORNELL, H. V., DAMSCHEN, E. I., JONATHAN DAVIES, T., GRYTNES, J.-A. et al. (2010). Niche conservatism as an emerging principle in ecology and conservation biology. *Ecol. Lett.* **13** 1310–1324.

WILLIAMS, R. J. and MARTINEZ, N. D. (2000). Simple rules yield complex food webs. *Nature* **404** 180–183.

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 241–256
https://doi.org/10.1214/19-AOAS1253

BAYESIAN FACTOR MODELS FOR PROBABILISTIC CAUSE OF DEATH ASSESSMENT WITH VERBAL AUTOPSIES

BY TSUYOSHI KUNIHAMA[1], ZEHANG RICHARD LI[2], SAMUEL J. CLARK[3] AND TYLER H. MCCORMICK[4]

[1]*Department of Economics, Kwansei Gakuin University, t.kunihama@kwansei.ac.jp*
[2]*Department of Biostatistics, Yale School of Public Health, zehang.li@yale.edu*
[3]*Department of Sociology, The Ohio State University, work@samclark.net*
[4]*Department of Statistics, Department of Sociology, University of Washington, tylermc@uw.edu*

The distribution of deaths by cause provides crucial information for public health planning, response and evaluation. About 60% of deaths globally are not registered or given a cause, limiting our ability to understand disease epidemiology. Verbal autopsy (VA) surveys are increasingly used in such settings to collect information on the signs, symptoms and medical history of people who have recently died. This article develops a novel Bayesian method for estimation of population distributions of deaths by cause using verbal autopsy data. The proposed approach is based on a multivariate probit model where associations among items in questionnaires are flexibly induced by latent factors. Using the Population Health Metrics Research Consortium labeled data that include both VA and medically certified causes of death, we assess performance of the proposed method. Further, we estimate important questionnaire items that are highly associated with causes of death. This framework provides insights that will simplify future data

REFERENCES

ABOUZAHR, C., CLELAND, J., COULLARE, F., MACFARLANE, S. B., NOTZON, F. C., SETEL, P., SZRETER, S., ANDERSON, R. N., BAWAH, A. A. et al. (2007). The way forward. *Lancet* **370** 1791–1799.

ARMINGER, G. and MUTHÉN, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis–Hastings algorithm. *Psychometrika* **63** 271–300.

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. MR2806429 https://doi.org/10.1093/biomet/asr013

BLOOMBERG, M. R. and BISHOP, J. (2015). Understanding death, extending life. *Lancet* **386** e18–e19.

BYASS, P., CHANDRAMOHAN, D., CLARK, S. J., D'AMBRUOSO, L., FOTTRELL, E., GRAHAM, W. J., HERBST, A. J., HODGSON, A., HOUNTON, S. et al. (2012). Strengthening standardised interpretation of verbal autopsy data: The new InterVA-4 tool. *Global Health Action* **5** 19281.

CRAMÉR, H. (1999). *Mathematical Methods of Statistics*. *Princeton Landmarks in Mathematics*. Princeton Univ. Press, Princeton, NJ. MR1816288

DE SAVIGNY, D., RILEY, I., CHANDRAMOHAN, D., ODHIAMBO, F., NICHOLS, E., NOTZON, S., ABOUZAHR, C., MITRA, R., COBOS MUÑOZ, D. et al. (2017). Integrating community-based verbal autopsy into civil registration and vital statistics (crvs): System-level considerations. *Global Health Action* **10** 1272882.

DOORNIK, J. A. (2007). Object-oriented matrix programming using Ox, 3rd ed. Timberlake Consultants, London, and www.doornik.com, Oxford.

DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Amer. Statist. Assoc.* **104** 1042–1051. MR2562004 https://doi.org/10.1198/jasa.2009.tm08439

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3235677

HOFF, P. D. (2009). *A First Course in Bayesian Statistical Methods*. *Springer Texts in Statistics*. Springer, New York. MR2648134 https://doi.org/10.1007/978-0-387-92407-6

HORTON, R. (2007). Counting for health. *Lancet* **370** 1526. https://doi.org/10.1016/S0140-6736(07)61418-4

Key words and phrases. Bayesian latent model, cause of death, conditional dependence, multivariate data, verbal autopsies, survey data.

JAMES, S. L., FLAXMAN, A. D. and MURRAY, C. J. (2011). Performance of the Tariff Method: Validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics* **9** 31.

JHA, P. (2014). Reliable direct measurement of causes of death in low-and middle-income countries. *BMC Medicine* **12** 19.

KING, G. and LU, Y. (2008). Verbal autopsy methods with multiple causes of death. *Statist. Sci.* **23** 78–91. MR2523943 https://doi.org/10.1214/07-STS247

KING, G., LU, Y. and SHIBUYA, K. (2010). Designing verbal autopsy studies. *Population Health Metrics* **8** 19.

KUNIHAMA, T., LI, Z. R., CLARK, S. J. and MCCORMICK, T. H. (2020). Supplement to "Bayesian factor models for probabilistic cause of death assessment with verbal autopsies." https://doi.org/10.1214/19-AOAS1253SUPPA, https://doi.org/10.1214/19-AOAS1253SUPPB.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR1925014 https://doi.org/10.1002/9781119013563

LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14** 41–67. MR2036762

LOPEZ, A. D. (1998). Counting the dead in China. *BMJ* **317** 1399–1400.

LOZANO, R., LOPEZ, A. D., ATKINSON, C., NAGHAVI, M., FLAXMAN, A. D. and MURRAY, C. J. (2011). Performance of physician-certified verbal autopsies: Multisite validation study using clinical diagnostic gold standards. *Population Health Metrics* **9** 32.

MAHER, D., BIRARO, S., HOSEGOOD, V., ISINGO, R., LUTALO, T., MUSHATI, P., NGWIRA, B., NYIRENDA, M., TODD, J. et al. (2010). Translating global health research aims into action: The example of the ALPHA network. *Tropical Medicine & International Health* **15** 321–328.

MATHERS, C. D., FAT, D. M., INOUE, M., RAO, C. and LOPEZ, A. D. (2005). Counting the dead and what they died from: An assessment of the global status of cause of death data. *Bulletin of the World Health Organization* **83** 171–177.

MCCORMICK, T. H., LI, Z. R., CALVERT, C., CRAMPIN, A. C., KAHN, K. and CLARK, S. J. (2016). Probabilistic cause-of-death assignment using verbal autopsies. *J. Amer. Statist. Assoc.* **111** 1036–1049. MR3561927 https://doi.org/10.1080/01621459.2016.1152191

MEALLI, F. and RUBIN, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **102** 995–1000. MR3431570 https://doi.org/10.1093/biomet/asv035

MIASNIKOF, P., GIANNAKEAS, V., GOMES, M., ALEKSANDROWICZ, L., SHESTOPALOFF, A. Y., ALAM, D., TOLLMAN, S., SAMARIKHALAJ, A. and JHA, P. (2015). Naive Bayes classifiers for verbal autopsies: Comparison to physician-based classification for 21,000 child and adult deaths. *BMC Medicine* **13** 286.

MIKKELSEN, L., PHILLIPS, D. E., ABOUZAHR, C., SETEL, P. W., DE SAVIGNY, D., LOZANO, R. and LOPEZ, A. D. (2015). A global assessment of civil registration and vital statistics systems: Monitoring data quality and progress. *Lancet* **386** 1395–1406.

MONTAGNA, S., TOKDAR, S. T., NEELON, B. and DUNSON, D. B. (2012). Bayesian latent factor regression for functional and longitudinal data. *Biometrics* **68** 1064–1073. MR3040013 https://doi.org/10.1111/j.1541-0420.2012.01788.x

MURRAY, C. J., LOPEZ, A. D., BLACK, R., AHUJA, R., ALI, S. M., BAQUI, A., DANDONA, L., DANTZER, E., DAS, V. et al. (2011). Population Health Metrics Research Consortium gold standard verbal autopsy validation study: Design, implementation, and development of analysis datasets. *Population Health Metrics* **9** 27.

NAVARRO, D. (2015). Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.5). Univ. Adelaide. Available at http://ua.edu.au/ccs/teaching/lsr.

NICHOLS, E. K., BYASS, P., CHANDRAMOHAN, D., CLARK, S. J., FLAXMAN, A. D., JAKOB, R., LEITAO, J., MAIRE, N., RAO, C. et al. (2018). The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0. PLoS. *Medicine* **15** e1002486.

PHILLIPS, D. E., ABOUZAHR, C., LOPEZ, A. D., MIKKELSEN, L., DE SAVIGNY, D., LOZANO, R., WILMOTH, J. and SETEL, P. W. (2015). Are well functioning civil registration and vital statistics systems associated with better health outcomes? *Lancet* **386** 1386–1394.

PHMRC (2013). Population Health Metrics Research Consortium gold standard verbal autopsy data 2005–2011. Available at http://ghdx.healthdata.org/record/population-health-metrics-research-consortium-gold-standard-verbal-autopsy-data-2005-2011.

R CORE TEAM (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.

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

RUZICKA, L. T. and LOPEZ, A. D. (1990). The use of cause-of-death statistics for health situation assessment: National and international experiences. *World Health Statistics Quarterly* **43** 249–258.

SANKOH, O. and BYASS, P. (2012). The INDEPTH Network: Filling vital gaps in global epidemiology. *Int. J. Epidemiol.* **41** 579–588.

SEAMAN, S., GALATI, J., JACKSON, D. and CARLIN, J. (2013). What is meant by "missing at random"? *Statist. Sci.* **28** 257–268. MR3112409 https://doi.org/10.1214/13-sts415

SERINA, P., RILEY, I., STEWART, A., JAMES, S. L., FLAXMAN, A. D., LOZANO, R., HERNANDEZ, B., MOONEY, M. D., LUNING, R. et al. (2015). Improving performance of the Tariff Method for assigning causes of death to verbal autopsies. *BMC Medicine* **13** 291.

SOLEMAN, N., CHANDRAMOHAN, D. and SHIBUYA, K. (2006). Verbal autopsy: Current practices and challenges. *Bulletin of the World Health Organization* **84** 239–245.

WORLD HEALTH ORGANIZATION (2012). Verbal Autopsy Standards: The 2012 WHO verbal autopsy instrument. Available at https://goo.gl/bQXXhG.

WORLD HEALTH ORGANIZATION (2017). Verbal Autopsy Standards: The 2016 WHO verbal autopsy instrument. Available at https://goo.gl/Hgt6es.

YANG, G., HU, J., RAO, K. Q., MA, J., RAO, C. and LOPEZ, A. D. (2005). Mortality registration and surveillance in China: History, current situation and challenges. *Population Health Metrics* **3** 3.

ZHOU, X., NAKAJIMA, J. and WEST, M. (2014). Bayesian forecasting and portfolio decisions using dynamic dependent sparse factor models. *Int. J. Forecast.* **30** 963–980.

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 257–275
https://doi.org/10.1214/19-AOAS1307

ESTIMATING THE HEALTH EFFECTS OF ENVIRONMENTAL MIXTURES USING BAYESIAN SEMIPARAMETRIC REGRESSION AND SPARSITY INDUCING PRIORS

BY JOSEPH ANTONELLI[1], MAITREYI MAZUMDAR[2,*], DAVID BELLINGER[2,†], DAVID CHRISTIANI[3,‡], ROBERT WRIGHT[4] AND BRENT COULL[3,§]

[1]*Department of Statistics, University of Florida, jantonelli@ufl.edu*
[2]*Harvard Medical School, *maitreyi.mazumdar@childrens.harvard.edu; †david.bellinger@childrens.harvard.edu*
[3]*Harvard T.H. Chan School of Public Health, ‡dchris@hsph.harvard.edu; §bcoull@hsph.harvard.edu*
[4]*Icahn School of Medicine at Mount Sinai, robert.wright@mssm.edu*

Humans are routinely exposed to mixtures of chemical and other environmental factors, making the quantification of health effects associated with environmental mixtures a critical goal for establishing environmental policy sufficiently protective of human health. The quantification of the effects of exposure to an environmental mixture poses several statistical challenges. It is often the case that exposure to multiple pollutants interact with each other to affect an outcome. Further, the exposure-response relationship between an outcome and some exposures, such as some metals, can exhibit complex, nonlinear forms, since some exposures can be beneficial and detrimental at different ranges of exposure. To estimate the health effects of complex mixtures, we propose a flexible Bayesian approach that allows exposures to interact with each other and have nonlinear relationships with the outcome. We induce sparsity using multivariate spike and slab priors to determine which exposures are associated with the outcome and which exposures interact with each other. The proposed approach is interpretable, as we can use the posterior probabilities of inclusion into the model to identify pollutants that interact with each other. We utilize our approach to study the impact of exposure to metals on child neurodevelopment in Bangladesh and find a nonlinear, interactive relationship between arsenic and manganese.

REFERENCES

ANTONELLI, J., MAZUMDAR, M., BELLINGER, D., CHRISTIANI, D., WRIGHT, R. and COULL, B. (2020). Supplement to "Estimating the health effects of environmental mixtures using Bayesian semiparametric regression and sparsity inducing priors." https://doi.org/10.1214/19-AOAS1307SUPP.

BREIMAN, L. (2001). Random forests, decision trees, and categorical predictors: The "absent levels" problem. *Mach. Learn.* **45** 5–32.

BAYLEY, N. (2006). *Bayley Scales of Infant and Toddler Development-Third Edition*: *Administration Manual*.

BENGIO, Y., DUCHARME, R., VINCENT, P. and JAUVIN, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.* **3** 1137–1155.

BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A LASSO for hierarchical interactions. *Ann. Statist.* **41** 1111–1141. MR3113805 https://doi.org/10.1214/13-AOS1096

BOBB, J. F., VALERI, L., CLAUS HENN, B., CHRISTIANI, D. C., WRIGHT, R. O., MAZUMDAR, M., GODLESKI, J. J. and COULL, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* **16** 493–508. MR3365442 https://doi.org/10.1093/biostatistics/kxu058

BOBB, J. F., HENN, B. C., VALERI, L. and COULL, B. A. (2018). Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environ. Health* **17** 67.

BRAUN, J. M. (2017). Early-life exposure to EDCs: Role in childhood obesity and neurodevelopment. *Nat. Rev. Endocrinol.* **13** 161–173. https://doi.org/10.1038/nrendo.2016.186

BRAUN, J. M., GENNINGS, C., HAUSER, R. and WEBSTER, T. F. (2016). What can epidemiological studies tell us about the impact of chemical mixtures on human health? *Environ. Health Perspect.* **124** A6–A9.

Key words and phrases. Environmental statistics, interaction model, Bayesian inference, spike and slab priors.

CARLIN, D. J., RIDER, C. V., WOYCHIK, R. and BIRNBAUM, L. S. (2013). Unraveling the health effects of environmental mixtures: An NIEHS priority. *Environ. Health Perspect.* **121** A6–A8. https://doi.org/10.1289/ehp.1206182

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172 https://doi.org/10.1214/09-AOAS285

CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge Univ. Press, Cambridge.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. *Texts in Statistical Science Series*. CRC Press/CRC, Boca Raton, FL. MR2027492

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. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GUO, F., WANG, X., FAN, K., BRODERICK, T. and DUNSON, D. B. (2016). Boosting variational inference. ArXiv Preprint. Available at arXiv:1611.05559.

HAHN, P. R., MURRAY, J. and CARVALHO, C. M. (2017). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. Confounding, and Heterogeneous Effects (October 5, 2017).

HAHN, P. R., CARVALHO, C. M., PUELZ, D. and HE, J. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal.* **13** 163–182. MR3737947 https://doi.org/10.1214/16-BA1044

HAO, N., FENG, Y. and ZHANG, H. H. (2014). Model selection for high dimensional quadratic regression via regularization **85721** 26.

HARLEY, K. G., BERGER, K., RAUCH, S., KOGUT, K., HENN, B. C., CALAFAT, A. M., HUEN, K., ESKENAZI, B. and HOLLAND, N. (2017). Association of prenatal urinary phthalate metabolite concentrations and childhood BMI and obesity. *Pediatr. Res.* **82** 405.

HENN, B. C., COULL, B. A. and WRIGHT, R. O. (2014). Chemical mixtures and children's health. *Curr. Opin. Pediatr.* **26** 223.

HENN, B. C., ETTINGER, A. S., SCHWARTZ, J., TÉLLEZ-ROJO, M. M., LAMADRID-FIGUEROA, H., HERNÁNDEZ-AVILA, M., SCHNAAS, L., AMARASIRIWARDENA, C., BELLINGER, D. C. et al. (2010). Early postnatal blood manganese levels and children's neurodevelopment. *Epidemiology* **21** 433.

KORTENKAMP, A., FAUST, M., SCHOLZE, M. and BACKHAUS, T. (2007). Low-level exposure to multiple chemicals: Reason for human health concerns? *Environ. Health Perspect.* **115** 106–114.

LAZAREVIC, N., BARNETT, A. G., SLY, P. D. and KNIBBS, L. D. (2019). Statistical methodology in studies of prenatal exposure to mixtures of endocrine-disrupting chemicals: A review of existing approaches and new alternatives. *Environ. Health Perspect.* **127** 026001.

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

MILLER, A. C., FOTI, N. and ADAMS, R. P. (2016). Variational boosting: Iteratively refining posterior approximations. ArXiv Preprint. Available at arXiv:1611.06585.

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. MR0997578

O'HAGAN, A. (1978). Curve fitting and optimal design for prediction. *J. Roy. Statist. Soc. Ser. B* **40** 1–42. MR0512140

QAMAR, S. and TOKDAR, S. (2014). Additive Gaussian process regression. ArXiv Preprint. Available at arXiv:1411.7009.

RADCHENKO, P. and JAMES, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *J. Amer. Statist. Assoc.* **105** 1541–1553. MR2796570 https://doi.org/10.1198/jasa.2010.tm10130

REICH, B. J., STORLIE, C. B. and BONDELL, H. D. (2009). Variable selection in Bayesian smoothing spline ANOVA models: Application to deterministic computer codes. *Technometrics* **51** 110–120. MR2668168 https://doi.org/10.1198/TECH.2009.0013

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

SCHEIPL, F., FAHRMEIR, L. and KNEIB, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *J. Amer. Statist. Assoc.* **107** 1518–1532. MR3036413 https://doi.org/10.1080/01621459.2012.737742

SHIVELY, T. S., KOHN, R. and WOOD, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior. *J. Amer. Statist. Assoc.* **94** 777–806. MR1723272 https://doi.org/10.2307/2669990

TAYLOR, K. W., JOUBERT, B. R., BRAUN, J. M., DILWORTH, C., GENNINGS, C., HAUSER, R., HEINDEL, J. J., RIDER, C. V., WEBSTER, T. F. et al. (2016). Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: Lessons from an innovative workshop. *Environ. Health Perspect.* **124** A227–A229. https://doi.org/10.1289/EHP547

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

VALERI, L., MAZUMDAR, M. M., BOBB, J. F., HENN, B. C., RODRIGUES, E., SHARIF, O. I., KILE, M. L., QUAMRUZZAMAN, Q., AFROZ, S. et al. (2017). The joint effect of prenatal exposure to metal mixtures on neurodevelopmental outcomes at 20–40 months of age: Evidence from rural Bangladesh. *Environ. Health Perspect.* **125**.

WATANABE, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11** 3571–3594. MR2756194

WOOD, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* **62** 1025–1036. MR2297673 https://doi.org/10.1111/j.1541-0420.2006.00574.x

WOOD, S., KOHN, R., SHIVELY, T. and JIANG, W. (2002). Model selection in spline nonparametric regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 119–139. MR1883129 https://doi.org/10.1111/1467-9868.00328

YAU, P., KOHN, R. and WOOD, S. (2003). Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression. *J. Comput. Graph. Statist.* **12** 23–54. MR1965210 https://doi.org/10.1198/1061860031301

ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497. MR2549566 https://doi.org/10.1214/07-AOS584

ZHOU, J., BHATTACHARYA, A., HERRING, A. H. and DUNSON, D. B. (2015). Bayesian factorizations of big sparse tensors. *J. Amer. Statist. Assoc.* **110** 1562–1576. MR3449055 https://doi.org/10.1080/01621459.2014.983233

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

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

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 276–298
https://doi.org/10.1214/19-AOAS1310

FEATURE SELECTION FOR GENERALIZED VARYING COEFFICIENT MIXED-EFFECT MODELS WITH APPLICATION TO OBESITY GWAS

BY WANGHUAN CHU[1], RUNZE LI[2], JINGYUAN LIU[3] AND MATTHEW REIMHERR[4]

[1]*Google Inc., dqchuwh@gmail.com*
[2]*Department of Statistics and the Methodology Center, Pennsylvania State University, rzli@psu.edu*
[3]*MOE Key Laboratory of Econometrics, Department of Statistics, School of Economics, Wang Yanan Institute for Studies in Economics, and Fujian Key Lab of Statistics, Xiamen University, jingyuan@xmu.edu.cn*
[4]*Department of Statistics, Pennsylvania State University, mreimherr@psu.edu*

Motivated by an empirical analysis of data from a genome-wide association study on obesity, measured by the body mass index (BMI), we propose a two-step gene-detection procedure for generalized varying coefficient mixed-effects models with ultrahigh dimensional covariates. The proposed procedure selects significant single nucleotide polymorphisms (SNPs) impacting the mean BMI trend, some of which have already been biologically proven to be "fat genes." The method also discovers SNPs that significantly influence the age-dependent variability of BMI. The proposed procedure takes into account individual variations of genetic effects and can also be directly applied to longitudinal data with continuous, binary or count responses. We employ Monte Carlo simulation studies to assess the performance of the proposed method and further carry out causal inference for the selected SNPs.

REFERENCES

ALLISON, D. B., KAPRIO, J., KORKEILA, M., KOSKENVUO, M., NEALE, M. C. and HAYAKAWA, K. (1996). The heritability of body mass index among an international sample of monozygotic twins reared apart. *Int. J. Obes.* **20** 501–506.

ASCHARD, H., ZAITLEN, N., TAMIMI, R. M., LINDSTRÖM, S. and KRAFT, P. (2013). A nonparametric test to detect quantitative trait loci where the phenotypic distribution differs by genotypes. *Genet. Epidemiol.* **37** 323–333.

BARBER, R. F., REIMHERR, M. and SCHILL, T. (2017). The function-on-scalar LASSO with applications to longitudinal GWAS. *Electron. J. Stat.* **11** 1351–1389. MR3635916 https://doi.org/10.1214/17-EJS1260

BECKER, J., NORA, D. B., GOMES, I., STRINGARI, F. F., SEITENSUS, R., PANOSSO, J. S. and EHLERS, J. A. C. (2002). An evaluation of gender, obesity, age and diabetes mellitus as risk factors for carpal tunnel syndrome. *Clin. Neurophysiol.* **113** 1429–1434.

BELL, C. G., WALLEY, A. J. and FROGUEL, P. (2005). The genetics of human obesity. *Nat. Rev. Genet.* **6** 221–234.

CHEN, H., QI, X., QIU, P. and ZHAO, J. (2015). Correlation between LSP1 polymorphisms and the susceptibility to breast cancer. *Int. J. Clin. Exp. Pathol.* **8** 5798–5802.

CHU, W., LI, R. and REIMHERR, M. (2016). Feature screening for time-varying coefficient models with ultrahigh-dimensional longitudinal data. *Ann. Appl. Stat.* **10** 596–617. MR3528353 https://doi.org/10.1214/16-AOAS912

CHU, W., LI, R., LIU, J. and REIMHERR, M. (2020). Supplement to "Feature selection for generalized varying coefficient mixed-effect models with application to obesity GWAS." https://doi.org/10.1214/19-AOAS1310SUPP.

DE JONGH, R. T., SERNÉ, E. H., IJZERMAN, R. G., DE VRIES, G. and STEHOUWER, C. D. (2004). Impaired microvascular function in obesity implications for obesity-associated microangiopathy, hypertension, and insulin resistance. *Circulation* **109** 2529–2535.

FALL, T. and INGELSSON, E. (2014). Genome-wide association studies of obesity and metabolic syndrome. *Mol. Cell. Endocrinol.* **382** 740–757. https://doi.org/10.1016/j.mce.2012.08.018

Key words and phrases. Genome-wide association study, mixed effects, ultrahigh dimensional longitudinal data, varying coefficient models.

FAN, J., FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557. MR2847969 https://doi.org/10.1198/jasa.2011.tm09779

FAN, J., GUO, S. and HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 37–65. MR2885839 https://doi.org/10.1111/j.1467-9868.2011.01005.x

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

FAN, J., MA, Y. and DAI, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *J. Amer. Statist. Assoc.* **109** 1270–1284. MR3265696 https://doi.org/10.1080/01621459.2013.879828

FURLOTTE, N. A. and ESKIN, E. (2015). Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. *Genetics* **200** 59–68. https://doi.org/10.1534/genetics.114.171447

GEILER-SAMEROTTE, K., BAUER, C., LI, S., ZIV, N., GRESHAM, D. and SIEGAL, M. (2013). The details in the distributions: Why and how to study phenotypic variability. *Curr. Opin. Biotechnol.* **24** 752–759.

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881

HERRERA, B. M., KEILDSON, S. and LINDGREN, C. M. (2011). Genetics and epigenetics of obesity. *Maturitas* **69** 41–49.

HILDEBRANDT, M. A., KOMAKI, R., LIAO, Z., GU, J., CHANG, J. Y., YE, Y., LU, C., STEWART, D. J., MINNA, J. D. et al. (2010). Genetic variants in inflammation-related genes are associated with radiation-induced toxicity following treatment for non-small cell lung cancer. *PLoS ONE* **5** e12402.

HSIEH, Y.-Y., CHANG, C.-C., HSU, C.-M., CHEN, S.-Y., LIN, W.-H. and TSAI, F.-J. (2011). Major histocompatibility complex class I chain-related gene polymorphisms: Associated with susceptibility to Kawasaki disease and coronary artery aneurysms. *Genet. Test. Mol. Biomark.* **15** 755–763.

HUNG, S.-I., CHUNG, W.-H., LIOU, L.-B., CHU, C.-C., LIN, M., HUANG, H.-P., LIN, Y.-L., LAN, J.-L., YANG, L.-C. et al. (2005). HLA-B* 5801 allele as a genetic marker for severe cutaneous adverse reactions caused by allopurinol. *Proc. Natl. Acad. Sci. USA* **102** 4134–4139.

JOHNSON, T., GAUNT, T. R., NEWHOUSE, S. J., PADMANABHAN, S., TOMASZEWSKI, M., KUMARI, M., MORRIS, R. W., TZOULAKI, I., O'BRIEN, E. T. et al. (2011). Blood pressure loci identified with a gene-centric array. *Am. J. Hum. Genet.* **89** 688–700. https://doi.org/10.1016/j.ajhg.2011.10.013

KAKLAMANI, V. G., SADIM, M., HSI, A., OFFIT, K., ODDOUX, C., OSTRER, H., AHSAN, H., PASCHE, B. and MANTZOROS, C. (2008). Variants of the adiponectin and adiponectin receptor 1 genes and breast cancer risk. *Cancer Res.* **68** 3178–3184.

KELLY, T., YANG, W., CHEN, C.-S., REYNOLDS, K. and HE, J. (2008). Global burden of obesity in 2005 and projections to 2030. *Int. J. Obes.* **32** 1431–1437.

LEE, K.-H., TSAI, W.-J., CHEN, Y.-W., YANG, W.-C., LEE, C.-Y., OU, S.-M., CHEN, Y.-T., CHIEN, C.-C., LEE, P.-C. et al. (2015). Genotype polymorphisms of genes regulating nitric oxide synthesis determine long-term arteriovenous fistula patency in male hemodialysis patients. *Ren. Fail.* **38** 1–10.

LIU, J., LI, R. and WU, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *J. Amer. Statist. Assoc.* **109** 266–274. MR3180562 https://doi.org/10.1080/01621459.2013.850086

LIU, J., ZHONG, W. and LI, R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Sci. China Math.* **58** 2033–2054. MR3400642 https://doi.org/10.1007/s11425-015-5062-9

MUNROE, P. B., BARNES, M. R. and CAULFIELD, M. J. (2013). Advances in blood pressure genomics. *Circ. Res.* **112** 1365–1379.

NAZIROĞLU, M., DIKICI, D. M. and DURSUN, S. Role of oxidative stress and Ca^{2+} signaling on molecular pathways of neuropathic pain in diabetes: Focus on TRP channels.

NGUYEN, J. D. U., LAMONTAGNE, M., COUTURE, C., CONTI, M., PARÉ, P. D., SIN, D. D., HOGG, J. C., NICKLE, D., POSTMA, D. S. et al. (2014). Susceptibility loci for lung cancer are associated with mRNA levels of nearby genes in the lung. *Carcinogenesis* **35** 2653–2659.

OGDEN, C. L., CARROLL, M. D., KIT, B. K. and FLEGAL, K. M. (2012). Prevalence of obesity in the United States, 2009–2010.

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

RAMACHANDRAPPA, S. and FAROOQI, I. S. (2011). Genetic approaches to understanding human obesity. *J. Clin. Invest.* **121** 2080–2086.

RAND, C. S. and KULDAU, J. M. (1990). The epidemiology of obesity and self-defined weight problem in the general population: Gender, race, age, and social class. *Int. J. Eat. Disord.* **9** 329–343.

RANKINEN, T., ZUBERI, A., CHAGNON, Y. C., WEISNAGEL, S. J., ARGYROPOULOS, G., WALTS, B., PÉRUSSE, L. and BOUCHARD, C. (2006). The human obesity gene map: The 2005 update. *Obesity* **14** 529–644. https://doi.org/10.1038/oby.2006.71

SOAVE, D., CORVOL, H., PANJWANI, N., GONG, J., LI, W., BOËLLE, P.-Y., DURIE, P. R., PATERSON, A. D., ROMMENS, J. M. et al. (2015). A joint location-scale test improves power to detect associated SNPs, gene sets, and pathways. *Am. J. Hum. Genet.* **97** 125–138.

SONG, R., YI, F. and ZOU, H. (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statist. Sinica* **24** 1735–1752. MR3308660

SUZUKI, T., IKARI, K., YANO, K., INOUE, E., TOYAMA, Y., TANIGUCHI, A., YAMANAKA, H. and MOMOHARA, S. (2013). PADI4 and HLA-DRB1 are genetic risks for radiographic progression in RA patients, independent of ACPA status: Results from the IORRA cohort study. *PLoS ONE* **8** e61045. https://doi.org/10.1371/journal.pone.0061045

TAPPER, W., HAMMOND, V., GERTY, S., ENNIS, S., SIMMONDS, P., COLLINS, A., ECCLES, D. and PROSPECTIVE STUDY OF OUTCOMES IN SPORADIC VERSUS HEREDITARY BREAST CANCER (POSH) STEERING GROUP (2008). The influence of genetic variation in 30 selected genes on the clinical characteristics of early onset breast cancer. *Breast Cancer Res.* **10** 1–10.

WANG, Y. and BEYDOUN, M. A. (2007). The obesity epidemic in the United States—gender, age, socioeconomic, racial/ethnic, and geographic characteristics: A systematic review and meta-regression analysis. *Epidemiol. Rev.* **29** 6–28. https://doi.org/10.1093/epirev/mxm007

WARDLE, J., CARNELL, S., HAWORTH, C. M. and PLOMIN, R. (2008). Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. *Am. J. Clin. Nutr.* **87** 398–404.

WEYER, C., FUNAHASHI, T., TANAKA, S., HOTTA, K., MATSUZAWA, Y., PRATLEY, R. E. and TATARANNI, P. A. (2001). Hypoadiponectinemia in obesity and type 2 diabetes: Close association with insulin resistance and hyperinsulinemia. *J. Clin. Endocrinol. Metab.* **86** 1930–1935.

WU, Z., CHEN, H., SUN, F., XU, J., ZHENG, W., LI, P., CHEN, S., SHEN, M., ZHANG, W. et al. (2013). PTPN2 rs1893217 single-nucleotide polymorphism is associated with risk of Behcet's disease in a Chinese Han population. *Clin. Exp. Rheumatol.* **32** S20–S26.

XIA, X., YANG, H. and LI, J. (2016). Feature screening for generalized varying coefficient models with application to dichotomous responses. *Comput. Statist. Data Anal.* **102** 85–97. MR3506984 https://doi.org/10.1016/j.csda.2016.04.008

YEOMANS, M. R. (2010). Alcohol, appetite and energy balance: Is alcohol intake a risk factor for obesity? *Physiol. Behav.* **100** 82–89.

ZHU, L.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849 https://doi.org/10.1198/jasa.2011.tm10563

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 299–338
https://doi.org/10.1214/19-AOAS1303

OPTIMAL ASSET ALLOCATION WITH MULTIVARIATE BAYESIAN DYNAMIC LINEAR MODELS

BY JARED D. FISHER[1], DAVIDE PETTENUZZO[2] AND CARLOS M. CARVALHO[3]

[1]*Department of Statistics, University of California, Berkeley, jared.fisher@berkeley.edu*
[2]*International Business School, Brandeis University, dpettenu@brandeis.edu*
[3]*McCombs School of Business, University of Texas at Austin, carlos.carvalho@mccombs.utexas.edu*

We introduce a fast, closed-form, simulation-free method to model and forecast multiple asset returns and employ it to investigate the optimal ensemble of features to include when jointly predicting monthly stock and bond excess returns. Our approach builds on the Bayesian dynamic linear models of West and Harrison (*Bayesian Forecasting and Dynamic Models* (1997) Springer), and it can objectively determine, through a fully automated procedure, both the optimal set of regressors to include in the predictive system and the degree to which the model coefficients, volatilities and covariances should vary over time. When applied to a portfolio of five stock and bond returns, we find that our method leads to large forecast gains, both in statistical and economic terms. In particular, we find that relative to a standard no-predictability benchmark, the optimal combination of predictors, stochastic volatility and time-varying covariances increases the annualized certainty equivalent returns of a leverage-constrained power utility investor by more than 500 basis points.

REFERENCES

ANG, A. and BEKAERT, G. (2007). Stock return predictability: Is it there? *Rev. Financ. Stud.* **20** 651–707. https://doi.org/10.1093/rfs/hhl021

BARNARD, J., MCCULLOCH, R. and MENG, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist. Sinica* **10** 1281–1311. MR1804544

BILLIO, M., CASARIN, R., RAVAZZOLO, F. and VAN DIJK, H. K. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *J. Econometrics* **177** 213–232. MR3118557 https://doi.org/10.1016/j.jeconom.2013.04.009

BOSSAERTS, P. and HILLION, P. (1999). Implementing statistical criteria to select return forecasting models: What do we learn? *Rev. Financ. Stud.* **12** 405–428. https://doi.org/10.1093/rfs/12.2.405

BRENNAN, M. J., SCHWARTZ, E. S. and LAGNADO, R. (1997). Strategic asset allocation. *J. Econom. Dynam. Control* **21** 1377–1403. MR1470286 https://doi.org/10.1016/S0165-1889(97)00031-6

CAMPBELL, J. Y., CHAN, Y. L. and VICEIRA, L. M. (2003). A multivariate model of strategic asset allocation. *J. Financ. Econ.* **67** 41–80. https://doi.org/10.1016/S0304-405X(02)00231-3

CAMPBELL, J. Y. and SHILLER, R. J. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Rev. Financ. Stud.* **1** 195–228. https://doi.org/10.1093/rfs/1.3.195

CARRIERO, A., CLARK, T. E. and MARCELLINO, M. (2019). Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *J. Econometrics* **212** 137–154. https://doi.org/10.1016/j.jeconom.2019.04.024

CHRISTOFFERSEN, P. F. and DIEBOLD, F. X. (1998). Cointegration and long-horizon forecasting. *J. Bus. Econom. Statist.* **16** 450–458. MR1650225 https://doi.org/10.2307/1392613

COCHRANE, J. H. and PIAZZESI, M. (2005). Bond risk premia. *Am. Econ. Rev.* **95** 138–160.

DANGL, T. and HALLING, M. (2012). Predictive regressions with time-varying coefficients. *J. Financ. Econ.* **106** 157–181. https://doi.org/10.1016/j.jfineco.2012.04.003

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

ENGLE, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *J. Bus. Econom. Statist.* **20** 339–350. MR1939905 https://doi.org/10.1198/073500102288618487

Key words and phrases. 62C10, 62P20, 91B28, 91B84.

FAMA, E. F. and BLISS, R. R. (1987). The information in long-maturity forward rates. *Am. Econ. Rev.* **77** 680–692.
FAMA, E. F. and SCHWERT, G. W. (1977). Asset returns and inflation. *J. Financ. Econ.* **5** 115–146. https://doi.org/10.1016/0304-405X(77)90014-9
FAN, J., FURGER, A. and XIU, D. (2016). Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. *J. Bus. Econom. Statist.* **34** 489–503. MR3547991 https://doi.org/10.1080/07350015.2015.1052458
GAO, X. and NARDARI, F. (2018). Do commodities add economic value in asset allocation? New evidence from time-varying moments. *J. Financ. Quant. Anal.* **53**.
GARGANO, A., PETTENUZZO, D. and TIMMERMANN, A. G. (2019). Bond return predictability: Economic value and links to the macroeconomy. *Manage. Sci.* **65** 508–540.
GELMAN, A. and HILL, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models. Analytical Methods for Social Research*. Cambridge Univ. Press, Cambridge. https://doi.org/10.1017/CBO9780511790942
GEWEKE, J. and AMISANO, G. (2011). Optimal prediction pools. *J. Econometrics* **164** 130–141. MR2821798 https://doi.org/10.1016/j.jeconom.2011.02.017
GRUBER, L. and WEST, M. (2016). GPU-accelerated Bayesian learning and forecasting in simultaneous graphical dynamic linear models. *Bayesian Anal.* **11** 125–149. MR3447094 https://doi.org/10.1214/15-BA946
GURKAYNAK, R. S., SACK, B. and WRIGHT, J. H. (2007). The U.S. Treasury yield curve: 1961 to the present. *J. Monet. Econ.* **54** 2291–2304. https://doi.org/10.1016/j.jmoneco.2007.06.029
JOHANNES, M., KORTEWEG, A. and POLSON, N. (2014). Sequential learning, predictability, and optimal portfolio returns. *J. Finance* **69** 611–644. https://doi.org/10.1111/jofi.12121
KIM, C.-J., MORLEY, J. C. and NELSON, C. R. (2005). The structural break in the equity premium. *J. Bus. Econom. Statist.* **23** 181–191. MR2157269 https://doi.org/10.1198/073500104000000352
KOOP, G., KOROBILIS, D. and PETTENUZZO, D. (2019). Bayesian compressed vector autoregressions. *J. Econometrics* **210** 135–154. MR3944767 https://doi.org/10.1016/j.jeconom.2018.11.009
LETTAU, M. and LUDVIGSON, S. (2001). Consumption, aggregate wealth, and expected stock returns. *J. Finance* **56** 815–849. https://doi.org/10.1111/0022-1082.00347
LETTAU, M. and VAN NIEUWERBURGH, S. (2008). Reconciling the return predictability evidence. *Rev. Financ. Stud.* **21** 1607–1652. https://doi.org/10.1093/rfs/hhm074
LEWELLEN, J. (2004). Predicting returns with financial ratios. *J. Financ. Econ.* **74** 209–235. https://doi.org/10.1016/j.jfineco.2002.11.002
LUDVIGSON, S. C. and NG, S. (2009). Macro factors in bond risk premia. *Rev. Financ. Stud.* **22** 5027–5067. https://doi.org/10.1093/rfs/hhp081
PASTOR, L. and STAMBAUGH, R. F. (2001). The equity premium and structural breaks. *J. Finance* **56** 1207–1239. https://doi.org/10.1111/0022-1082.00365
PAYE, B. S. and TIMMERMANN, A. (2006). Instability of return prediction models. *J. Empir. Finance* **13** 274–315.
PETTENUZZO, D. and RAVAZZOLO, F. (2016). Optimal portfolio choice under decision-based model combinations. *J. Appl. Econometrics* **31** 1312–1332. MR3580902 https://doi.org/10.1002/jae.2502
PETTENUZZO, D. and TIMMERMANN, A. (2011). Predictability of stock returns and asset allocation under structural breaks. *J. Econometrics* **164** 60–78. MR2821794 https://doi.org/10.1016/j.jeconom.2011.02.019
PETTENUZZO, D., TIMMERMANN, A. and VALKANOV, R. (2014). Forecasting stock returns under economic constraints. *J. Financ. Econ.* **114** 517–553.
PRIMICERI, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *Rev. Econ. Stud.* **72** 821–852. MR2148143 https://doi.org/10.1111/j.1467-937X.2005.00353.x
RAPACH, D. E., STRAUSS, J. K. and ZHOU, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Rev. Financ. Stud.* **23** 821–862. https://doi.org/10.1093/rfs/hhp063
THORNTON, D. L. and VALENTE, G. (2012). Out-of-sample predictions of bond excess returns and forward rates: An asset allocation perspective. *Rev. Financ. Stud.* **25** 3141–3168.
VICEIRA, L. (1997). Testing for structural change in the predictability of asset returns. Unpublished manuscript.
WACHTER, J. A. and WARUSAWITHARANA, M. (2009). Predictable returns and asset allocation: Should a skeptical investor time the market? *J. Econometrics* **148** 162–178. MR2500654 https://doi.org/10.1016/j.jeconom.2008.10.009
WELCH, I. and GOYAL, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Rev. Financ. Stud.* **21** 1455–1508.
WEST, M. and HARRISON, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1482232
ZHAO, Z. Y., XIE, M. and WEST, M. (2016). Dynamic dependence networks: Financial time series forecasting and portfolio decisions. *Appl. Stoch. Models Bus. Ind.* **32** 311–332. MR3518020 https://doi.org/10.1002/asmb.2161

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 339–356
https://doi.org/10.1214/19-AOAS1309

MODELING WILDFIRE IGNITION ORIGINS IN SOUTHERN CALIFORNIA USING LINEAR NETWORK POINT PROCESSES

BY MEDHA UPPALA* AND MARK S. HANDCOCK†

*Department of Statistics, University of California Los Angeles, *umedha@ucla.edu; †handcock@ucla.edu*

This paper focuses on spatial and temporal modeling of point processes on linear networks. Point processes on linear networks can simply be defined as point events occurring on or near line segment network structures embedded in a certain space. A separable modeling framework is introduced that posits separate *formation* and *dissolution* models of point processes on linear networks over time. While the model was inspired by spider web building activity in brick mortar lines, the focus is on modeling wildfire ignition origins near road networks over a span of 14 years. As most wildfires in California have human-related origins, modeling the origin locations with respect to the road network provides insight into how human, vehicular and structural densities affect ignition occurrence. Model results show that roads that traverse different types of regions such as residential, interface and wildland regions have higher ignition intensities compared to roads that only exist in each of the mentioned region types.

REFERENCES

ANG, Q. W., BADDELEY, A. and NAIR, G. (2012). Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scand. J. Stat.* **39** 591–617. MR3000837 https://doi.org/10.1111/j.1467-9469.2011.00752.x

BADDELEY, A., JAMMALAMADAKA, A. and NAIR, G. (2014). Multitype point process analysis of spines on the dendrite network of a neuron. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 673–694. MR3269407 https://doi.org/10.1111/rssc.12054

BADDELEY, A. and TURNER, R. (2000). Practical maximum pseudolikelihood for spatial point patterns (with discussion). *Aust. N. Z. J. Stat.* **42** 283–322. MR1794056 https://doi.org/10.1111/1467-842X.00128

BADDELEY, A., TURNER, R., MØLLER, J. and HAZELTON, M. (2005). Residual analysis for spatial point processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 617–666. MR2210685 https://doi.org/10.1111/j.1467-9868.2005.00519.x

BALCH, J. K., BRADLEY, B. A., ABATZOGLOU, J. T., NAGY, R. C., FUSCO, E. J. and MAHOOD, A. L. (2017). Human-started wildfires expand the fire niche across the United States. *Proc. Natl. Acad. Sci. USA* **114** 2946–2951.

BERMAN, M. and TURNER, R. (1992). Approximating point process likelihoods with GLIM. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **41** 31–38.

BESAG, J. and DIGGLE, P. J. (1977). Simple Monte Carlo tests for spatial pattern. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **26** 327–333.

CLEMENTS, R. A., SCHOENBERG, F. P. and VEEN, A. (2012). Evaluation of space-time point process models using super-thinning. *Environmetrics* **23** 606–616. MR3020078 https://doi.org/10.1002/env.2168

HALL, P. (1985). Resampling a coverage pattern. *Stochastic Process. Appl.* **20** 231–246. MR0808159 https://doi.org/10.1016/0304-4149(85)90212-1

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. CRC Press, London. MR1082147

HERING, A. S. and BAIR, S. (2014). Characterizing spatial and chronological target selection of serial offenders. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 123–140. MR3148272 https://doi.org/10.1111/rssc.12029

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

KÜNSCH, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241. MR1015147 https://doi.org/10.1214/aos/1176347265

Key words and phrases. Point processes, linear network, spatiotemporal modeling, pseudolikelihood, Berman–Turner method, spider webs, wildfires, ignition origins, road networks.

LIU, R. Y. and SINGH, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the Limits of Bootstrap* (*East Lansing*, *MI*, 1990) (R. Lepage and L. Billard eds). 225–248. Wiley, New York. MR1197787

MANN, M. L., BERCK, P., MORITZ, M. A., BATLLORI, E., BALDWIN, J. G., GATELY, C. K. and CAMERON, D. R. (2014). Modeling residential development in California from 2000 to 2050: Integrating wildfire risk, wildland and agricultural encroachment. *Land Use Policy* **41** 438–452.

OKABE, A. and OKUNUKI, K. I. (2001). A computational method for estimating the demand of retail stores on a street network and its implementation in GIS. *Trans*. *GIS* **5** 209–220.

OKABE, A. and SUGIHARA, K. (2012). *Spatial Analysis Along Networks*: *Statistical and Computational Methods*. Wiley, Chichester.

OKABE, A. and YAMADA, I. (2001). The K-function method on a network and its computational implementation. *Geogr. Anal*. **33**.

PAPANGELOU, F. (1973/74). The conditional intensity of general point processes and an application to line processes. *Z. Wahrsch. Verw. Gebiete* **28** 207–226. MR0373000 https://doi.org/10.1007/BF00533242

RADELOFF, V. C., HELMERS, D. P., KRAMER, H. A., MOCKRIN, M. H., ALEXANDRE, P. M., BAR MASSADA, A., BUTSIC, V., HAWBAKER, T. J., MARTINUZZI, S. et al. (2017). The 1990–2010 wildland-urban interface of the conterminous United States—geospatial data [2nd ed.].

RADELOFF, V. C., HELMERS, D. P., KRAMER, H. A., MOCKRIN, M. H., ALEXANDRE, P. M., BAR-MASSADA, A., BUTSIC, V., HAWBAKER, T. J., MARTINUZZI, S. et al. (2018). Rapid growth of the US wildland-urban interface raises wildfire risk. *Proc. Natl. Acad. Sci. USA* **115** 3314–3319. https://doi.org/10.1073/pnas.1718850115

SCHLESINGER, T. and EUGSTER, M. J. A. (2010). osmar: OpenStreetMap and R. *R J*.

SHORT, K. C. (2015). Spatial wildfire occurrence data for the United States, 1992–2013 [FPA_FOD_20150323] (3rd ed.).

SPOONER, P. G., LUNT, I. D., OKABE, A. and SHIODE, S. (2004). Spatial analysis of roadside Acacia populations on a road network using the network K-function. *Landsc. Ecol*. **19** 491–499.

SYPHARD, A. D. and KEELEY, J. E. (2015). Location, timing and extent of wildfire vary by cause of ignition. *Int. J. Wildland Fire* **24** 37–47.

XU, H. and SCHOENBERG, F. P. (2011). Point process modeling of wildfire hazard in Los Angeles County, California. *Ann. Appl. Stat*. **5** 684–704. MR2840171 https://doi.org/10.1214/10-AOAS401

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 357–380
https://doi.org/10.1214/19-AOAS1299

REGRESSION FOR COPULA-LINKED COMPOUND DISTRIBUTIONS WITH APPLICATIONS IN MODELING AGGREGATE INSURANCE CLAIMS

BY PENG SHI[1] AND ZIFENG ZHAO[2]

[1]*Department of Risk and Insurance, Wisconsin School of Business, University of Wisconsin—Madison, pshi@bus.wisc.edu*
[2]*Department of Information Technology, Analytics, and Operations, Mendoza College of Business, University of Notre Dame, zzhao2@nd.edu*

In actuarial research a task of particular interest and importance is to predict the loss cost for individual risks so that informative decisions are made in various insurance operations such as underwriting, ratemaking and capital management. The loss cost is typically viewed to follow a compound distribution where the summation of the severity variables is stopped by the frequency variable. A challenging issue in modeling such outcomes is to accommodate the potential dependence between the number of claims and the size of each individual claim. In this article we introduce a novel regression framework for compound distributions that uses a copula to accommodate the association between the frequency and the severity variables and, thus, allows for arbitrary dependence between the two components. We further show that the new model is very flexible and is easily modified to account for incomplete data due to censoring or truncation. The flexibility of the proposed model is illustrated using both simulated and real data sets. In the analysis of granular claims data from property insurance, we find substantive negative relationship between the number and the size of insurance claims. In addition, we demonstrate that ignoring the frequency-severity association could lead to biased decision-making in insurance operations.

REFERENCES

ACAR, E. F., CRAIU, R. V. and YAO, F. (2011). Dependence calibration in conditional copulas: A nonparametric approach. *Biometrics* **67** 445–453. MR2829013 https://doi.org/10.1111/j.1541-0420.2010.01472.x

ALBRECHER, H., BEIRLANT, J. and TEUGELS, J. L. (2017). *Reinsurance*: *Actuarial and Statistical Aspects*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3791478

ANTONIO, K. and BEIRLANT, J. (2008). Issues in claims reserving and credibility: A semiparametric approach with mixed models. *J. Risk Insur.* **75** 643–676.

ARIBARG, A., PIETERS, R. and WEDEL, M. (2010). Raising the BAR: Bias adjustment of recognition tests in advertising. *J. Mark. Res.* **47** 387–400.

CASTRO-CAMILO, D., DE CARVALHO, M. and WADSWORTH, J. (2018). Time-varying extreme value dependence with application to leading European stock markets. *Ann. Appl. Stat.* **12** 283–309. MR3773394 https://doi.org/10.1214/17-AOAS1089

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

CZADO, C., KASTENMEIER, R., BRECHMANN, E. C. and MIN, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scand. Actuar. J.* **4** 278–305. MR3010604 https://doi.org/10.1080/03461238.2010.546147

FERMANIAN, J.-D. and WEGKAMP, M. H. (2012). Time-dependent copulas. *J. Multivariate Anal.* **110** 19–29. MR2927507 https://doi.org/10.1016/j.jmva.2012.02.018

FREES, E. W. (2014). Frequency and severity models. In *Predictive Modeling Applications in Actuarial Science*, *Volume I*: *Predictive Modeling Techniques* (E. W. Frees, G. Meyers and R. A. Derrig, eds.) 138–166. Cambridge Univ. Press, Cambridge.

FREES, E. W. (2015). Analytics of insurance markets. *Annu. Rev. Financ. Econ.* **7** 253–277.

FREES, E. W., GAO, J. and ROSENBERG, M. A. (2011). Predicting the frequency and amount of health care expenditures. *N. Am. Actuar. J.* **15** 377–392. MR2869681 https://doi.org/10.1080/10920277.2011.10597626

Key words and phrases. Insurance claim data, compound distribution, copula regression, incomplete data, two-part model.

FREES, E. W., LEE, G. and YANG, L. (2016). Multivariate frequency-severity regression models in insurance. *Risks* **4** 4.

FREES, E. W., MEYERS, G. and CUMMINGS, A. D. (2011). Summarizing insurance scores using a Gini index. *J. Amer. Statist. Assoc.* **106** 1085–1098. MR2894766 https://doi.org/10.1198/jasa.2011.tm10506

GARRIDO, J., GENEST, C. and SCHULZ, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance Math. Econom.* **70** 205–215. MR3543046 https://doi.org/10.1016/j.insmatheco.2016.06.006

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548 https://doi.org/10.1198/016214506000001437

JOE, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *J. Multivariate Anal.* **94** 401–419. MR2167922 https://doi.org/10.1016/j.jmva.2004.06.003

JOE, H. (2015). *Dependence Modeling with Copulas*. *Monographs on Statistics and Applied Probability* **134**. CRC Press, Boca Raton, FL. MR3328438

JOHNSON, N. L., KEMP, A. W. and KOTZ, S. (2005). *Univariate Discrete Distributions*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR2163227 https://doi.org/10.1002/0471715816

JØRGENSEN, B. (1987). Exponential dispersion models. *J. Roy. Statist. Soc. Ser. B* **49** 127–162. MR0905186

JØRGENSEN, B. and PAES DE SOUZA, M. C. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scand. Actuar. J.* **1** 69–93. MR1286486 https://doi.org/10.1080/03461238.1994.10413930

KARLIS, D. and XEKALAKI, E. (2005). Mixed Poisson distributions. *Int. Stat. Rev.* **73** 35–58.

KLUGMAN, S. A., PANJER, H. H. and WILLMOT, G. E. (2012). *Loss Models*: *From Data to Decisions*, 4th ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3222004

KRÄMER, N., BRECHMANN, E. C., SILVESTRINI, D. and CZADO, C. (2013). Total loss estimation using copula-based regression models. *Insurance Math. Econom.* **53** 829–839. MR3130478 https://doi.org/10.1016/j.insmatheco.2013.09.003

LIN, X. S. (2014). Compound distributions. *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat04411.

LIU, H. and WANG, R. (2017). Collective risk models with dependence uncertainty. *Astin Bull.* **47** 361–389. MR3654415 https://doi.org/10.1017/asb.2017.4

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

NELSEN, R. B. (2006). *An Introduction to Copulas*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2197664 https://doi.org/10.1007/s11229-005-3715-x

NEWEY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, *Vol. IV*. *Handbooks in Econom.* **2** 2111–2245. North-Holland, Amsterdam. MR1315971

OLSEN, M. K. and SCHAFER, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *J. Amer. Statist. Assoc.* **96** 730–745. MR1946438 https://doi.org/10.1198/016214501753168389

PANJER, H. H. (2006). *Operational Risk*: *Modeling Analytics*. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR2244881 https://doi.org/10.1002/0470051310

PATTON, A. J. (2006). Modelling asymmetric exchange rate dependence. *Internat. Econom. Rev.* **47** 527–556. MR2216591 https://doi.org/10.1111/j.1468-2354.2006.00387.x

PIGEON, M., ANTONIO, K. and DENUIT, M. (2014). Individual loss reserving using paid-incurred data. *Insurance Math. Econom.* **58** 121–131. MR3257343 https://doi.org/10.1016/j.insmatheco.2014.06.012

SHEVCHENKO, P. V. (2010). Calculation of aggregate loss distributions. *J. Oper. Risk* **5** 3.

SHI, P. (2014). Fat-tailed regression models. In *Predictive Modeling Applications in Actuarial Science* (E. W. Frees, G. Meyers and R. A. Derrig, eds.) Cambridge Univ. Press, Cambridge.

SHI, P., FENG, X. and BOUCHER, J.-P. (2016). Multilevel modeling of insurance claims using copulas. *Ann. Appl. Stat.* **10** 834–863. MR3528362 https://doi.org/10.1214/16-AOAS914

SHI, P. and YANG, L. (2018). Pair copula constructions for insurance experience rating. *J. Amer. Statist. Assoc.* **113** 122–133. MR3803444 https://doi.org/10.1080/01621459.2017.1330692

SHI, P. and ZHAO, Z. (2020). Supplement to "Regression for copula-linked compound distributions with applications in modeling aggregate insurance claims." https://doi.org/10.1214/19-AOAS1299SUPP.

SILVA, J. M. C. and WINDMEIJER, F. (2001). Two-part multiple spell models for health care demand. *J. Econometrics* **104** 67–89. MR1864227 https://doi.org/10.1016/S0304-4076(01)00059-8

SMITH, M. D. (2003). Modelling sample selection using Archimedean copulas. *Econom. J.* **6** 99–123. MR1992394 https://doi.org/10.1111/1368-423X.00101

SMITHSON, M. and SHOU, Y. (2014). Randomly stopped sums: Models and psychological applications. *Front. Psychol.* **5** 1–11.

SMYTH, G. K. and JØRGENSEN, B. (2002). Fitting Tweedie's compound Poisson model to insurance claims data: Dispersion modelling. *Astin Bull.* **32** 143–157. MR1930491 https://doi.org/10.2143/AST.32.1.1020

TELLIS, G. J. (1988). Advertising exposure, loyalty, and brand purchase: A two-stage model of choice. *J. Mark. Res.* **25** 134–144.

TWEEDIE, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics*: *Applications and New Directions* (*Calcutta*, 1981) 579–604. Indian Statist. Inst., Calcutta. MR0786162

VERAVERBEKE, N., OMELKA, M. and GIJBELS, I. (2011). Estimation of a conditional copula and association measures. *Scand. J. Stat.* **38** 766–780. MR2859749 https://doi.org/10.1111/j.1467-9469.2011.00744.x

WÜTHRICH, M. V. and MERZ, M. (2008). *Stochastic Claims Reserving Methods in Insurance*. Wiley, New York.

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 381–408
https://doi.org/10.1214/19-AOAS1306

ESTIMATING AND FORECASTING THE SMOKING-ATTRIBUTABLE MORTALITY FRACTION FOR BOTH GENDERS JOINTLY IN OVER 60 COUNTRIES

BY YICHENG LI[*] AND ADRIAN E. RAFTERY[†]

Department of Statistics, University of Washington, [*]yl83@uw.edu; [†]raftery@uw.edu

Smoking is one of the leading preventable threats to human health and a major risk factor for lung cancer, upper aerodigestive cancer and chronic obstructive pulmonary disease. Estimating and forecasting the smoking attributable fraction (SAF) of mortality can yield insights into smoking epidemics and also provide a basis for more accurate mortality and life expectancy projection. Peto et al. (*Lancet* **339** (1992) 1268–1278) proposed a method to estimate the SAF using the lung cancer mortality rate as an indicator of exposure to smoking in the population of interest. Here, we use the same method to estimate the all-age SAF (ASAF) for both genders for over 60 countries. We document a strong and cross-nationally consistent pattern of the evolution of the SAF over time. We use this as the basis for a new Bayesian hierarchical model to project future male and female ASAF from over 60 countries simultaneously. This gives forecasts as well as predictive distributions that can be used to find uncertainty intervals for any quantity of interest. We assess the model using out-of-sample predictive validation and find that it provides good forecasts and well-calibrated forecast intervals, comparing favorably with other methods.

REFERENCES

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.

AU, J. S., MANG, O. W., FOO, W. and LAW, S. C. (2004). Time trends of lung cancer incidence by histologic types and smoking prevalence in Hong Kong 1983–2000. *Lung Cancer* **45** 143–152.

BONGAARTS, J. (2006). How long will we live? *Population and Development Review* **32** 605–628.

BONGAARTS, J. (2014). Trends in causes of death in low-mortality countries: Implications for mortality projections. *Population and Development Review* **40** 189–212.

BRITTON, J. (2017). Death, disease, and tobacco. *Lancet* **389** 1861–1862.

BURNS, D. M., LEE, L., SHEN, L. Z., GILPIN, E., TOLLEY, H. D., VAUGHN, J., SHANKS, T. G. et al. (1997). Cigarette smoking behavior in the United States. *Changes in Cigarette-related Disease Risks and Their Implication for Prevention and Control. Smoking and Tobacco Control Monograph* **8** 13–42.

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. et al. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76**.

CENTERS FOR DISEASE CONTROL (2019). What are the risk factors for lung cancer? Last accessed: Oct. 19, 2019. Available at https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm.

CHEN, Z., PETO, R., ZHOU, M., IONA, A., SMITH, M., YANG, L., GUO, Y., CHEN, Y., BIAN, Z. et al. (2015). Contrasting male and female trends in tobacco-attributed mortality in China: Evidence from successive nationwide prospective cohort studies. *Lancet* **386** 1447–1456.

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

EZZATI, M. and LOPEZ, A. D. (2003). Estimates of global mortality attributable to smoking in 2000. *Lancet* **362** 847–852.

Key words and phrases. Smoking attributable fraction, Peto–Lopez method, Bayesian hierarchical model, double logistic curve, probabilislic projection.

EZZATI, M. and LOPEZ, A. D. (2004). Regional, disease specific patterns of smoking-attributable mortality in 2000. *Tobacco Control* **13** 388–395.
FENELON, A. and PRESTON, S. H. (2012). Estimating smoking-attributable mortality in the United States. *Demography* **49** 797–818. https://doi.org/10.1007/s13524-012-0108-x
FOKAS, N. (2007). Growth functions, social diffusion, and social change. *Review of Sociology* **13** 5–30.
GAJALAKSHMI, V., PETO, R., KANAKA, T. S. and JHA, P. (2003). Smoking and mortality from tuberculosis and other diseases in India: Retrospective study of 43000 adult male deaths and 35000 controls. *Lancet* **362** 507–515.
GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548 https://doi.org/10.1198/016214506000001437
GRÜBLER, A., NAKIĆENOVIĆ, N. and VICTOR, D. G. (1999). Dynamics of energy technologies and global change. *Energy Policy* **27** 247–280.
HARVEY, A. C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge Univ. Press, Cambridge.
HEAD, G. A. and MCCARTY, R. (1987). Vagal and sympathetic components of the heart rate range and gain of the baroreceptor-heart rate reflex in conscious rats. *Journal of the Autonomic Nervous System* **21** 203–213.
HEAD, G. A., LUKOSHKOVA, E. V., MAYOROV, D. N. and VAN DEN BUUSE, M. (2004). Non-symmetrical double-logistic analysis of 24-h blood pressure recordings in normotensive and hypertensive rats. *Journal of Hypertension* **22** 2075–2085.
HOFFMAN, M. D. and GELMAN, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15** 1593–1623. MR3214779
ISLAMI, F., TORRE, L. A. and JEMAL, A. (2015). Global trends of lung cancer mortality and smoking prevalence. *Transl. Lung Cancer Res.* **4** 327–338. https://doi.org/10.3978/j.issn.2218-6751.2015.08.04
JANSSEN, F., VAN WISSEN, L. J. and KUNST, A. E. (2013). Including the smoking epidemic in internationally coherent mortality projections. *Demography* **50** 1341–1362.
KONG, K. A., JUNG-CHOI, K.-H., LIM, D., LEE, H. A., LEE, W. K., BAIK, S. J., PARK, S. H. and PARK, H. (2016). Comparison of prevalence-and smoking impact ratio-based methods of estimating smoking-attributable fractions of deaths. *J. Epidemiol.* **26** 145–154.
KUCHARAVY, D. and DE GUIO, R. (2011). Logistic substitution model and technological forecasting. *Procedia Engineering* **9** 402–416.
LAM, T., HO, S., HEDLEY, A., MAK, K. and PETO, R. (2001). Mortality and smoking in Hong Kong: Case-control study of all adult deaths in 1998. *BMJ* **323** 361.
LARISCY, J. T., HUMMER, R. A. and ROGERS, R. G. (2018). Cigarette smoking and all-cause and cause-specific adult mortality in the United States. *Demography* **55** 1855–1885. https://doi.org/10.1007/s13524-018-0707-2
LEVIN, M. L. (1953). The occurrence of lung cancer in man. *Acta Unio. Int. Contra. Cancrum.* **9** 531–541.
LI, Y. and RAFTERY, A. E (2020). Supplement to "Estimating and forecasting the smoking-attributable mortality fraction for both genders jointly in over 60 countries." https://doi.org/10.1214/19-AOAS1306SUPP.
LIU, B.-Q., PETO, R., CHEN, Z.-M., BOREHAM, J., WU, Y.-P., LI, J.-Y., CAMPBELL, T. C. and CHEN, J.-S. (1998). Emerging tobacco hazards in China: 1. Retrospective proportional mortality study of one million deaths. *BMJ* **317** 1411–1422.
LUO, L. (2013). Assessing validity and application scope of the intrinsic estimator approach to the age-period-cohort problem (with discussion). *Demography* **50** 1945–1988.
LUO, Q., YU, X. Q., WADE, S., CARUANA, M., PESOLA, F., CANFELL, K. and O'CONNELL, D. L. (2018). Lung cancer mortality in Australia: Projected outcomes to 2040. *Lung Cancer* **125** 68–76. https://doi.org/10.1016/j.lungcan.2018.09.001
MA, J., SIEGEL, R. L., JACOBS, E. J. and JEMAL, A. (2018). Smoking-attributable mortality by state in 2014, US. *Am. J. Prev. Med.* **54** 661–670.
MACKENBACH, J. P., HUISMAN, M., ANDERSEN, O., BOPP, M., BORGAN, J.-K., BORRELL, C., COSTA, G., DEBOOSERE, P., DONKIN, A. et al. (2004). Inequalities in lung cancer mortality by the educational level in 10 European populations. *European Journal of Cancer* **40** 126–135.
MARCHETTI, C., MEYER, P. S. and AUSUBEL, J. H. (1996). Human population dynamics revisited with the logistic model: How much can be modeled and predicted? *Technol. Forecast. Soc. Change* **52** 1–30. https://doi.org/10.1016/0040-1625(96)00001-7
MEHTA, N. and PRESTON, S. H. (2012). Continued increases in the relative risk of death from smoking. *American Journal of Public Health* **102** 2181–2186.
MISHRA, S., JOSEPH, R. A., GUPTA, P. C., PEZZACK, B., RAM, F., SINHA, D. N., DIKSHIT, R., PATRA, J. and JHA, P. (2016). Trends in bidi and cigarette smoking in India from 1998 to 2015, by age, gender and education. *BMJ Global Health* **1** e000005.
MONS, U. and BRENNER, H. (2017). Demographic ageing and the evolution of smoking-attributable mortality: The example of Germany. *Tob. Control* **26** 455–457. https://doi.org/10.1136/tobaccocontrol-2016-053008

MUSZYŃSKA, M. M., FIHEL, A. and JANSSEN, F. (2014). Role of smoking in regional variation in mortality in Poland. *Addiction* **109** 1931–1941.
NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 113–162. CRC Press, Boca Raton, FL. MR2858447
NG, M., FREEMAN, M. K., FLEMING, T. D., ROBINSON, M., DWYER-LINDGREN, L., THOMSON, B., WOLLUM, A., SANMAN, E., WULF, S. et al. (2014). Smoking prevalence and cigarette consumption in 187 countries, 1980–2012. *Journal of the American Medical* **311** 183–192.
NIU, S.-R., YANG, G.-H., CHEN, Z.-M., WANG, J.-L., WANG, G.-H., HE, X.-Z., SCHOEPFF, H., BOREHAM, J., PAN, H.-C. et al. (1998). Emerging tobacco hazards in China: 2. Early mortality results from a prospective study. *BMJ* **317** 1423–1424.
PAMPEL, F. (2005). Forecasting sex differences in mortality in high income nations: The contribution of smoking. *Demogr. Res*. **13** 455–484. https://doi.org/10.4054/DemRes.2005.13.18
PAMPEL, F. C. (2006). Global patterns and determinants of sex differences in smoking. *Int. J. Comp. Sociol*. **47** 466–487. https://doi.org/10.1177/0020715206070267
PARASCANDOLA, M. and XIAO, L. (2019). Tobacco and the lung cancer epidemic in China. *Transl. Lung Cancer Res*. **8** S21–S30. https://doi.org/10.21037/tlcr.2019.03.12
PÉREZ-RÍOS, M. and MONTES, A. (2008). Methodologies used to estimate tobacco-attributable mortality: A review. *BMC Public Health* **8** 22.
PETERS, F., MACKENBACH, J. and NUSSELDER, W. (2016). Do life expectancy projections need to account for the impact of smoking. *Netspar. Design Papers* **2016** 1–54.
PETO, R., BOREHAM, J., LOPEZ, A. D., THUN, M. and HEATH, C. (1992). Mortality from tobacco in developed countries: Indirect estimation from national vital statistics. *Lancet* **339** 1268–1278.
PETO, R., LOPEZ, A. D., BOREHAM, J., THUN, M. and HEATH, C. (1994). Mortality from smoking in developed countries 1950–2000. Indirect estimates from national statistics.
PETO, R., LOPEZ, A. D., BOREHAM, J. and THUN, M. (2006). Mortality from smoking in developed countries. *Population* **673290** 300245.
PRESTON, S. H., GLEI, D. A. and WILMOTH, J. R. (2009). A new method for estimating smoking-attributable mortality in high-income countries. *Int. J. Epidemiol*. **39** 430–438.
PRESTON, S. H., GLEI, D. A. and WILMOTH, J. R. (2011). Contribution of smoking to international differences in life expectancy. In *International Differences in Mortality at Older Ages*: *Dimensions and Sources* (E. M. Crimmins, S. H. Preston and B. Cohen, eds.) 105–31. The National Academies Press, Washington, DC.
PRESTON, S. H. and WANG, H. (2006). Sex mortality differences in the United States: The role of cohort smoking patterns. *Demography* **43** 631–646.
RAFTERY, A. E., CHUNN, J. L., GERLAND, P. and SEVČÍKOVÁ, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography* **50** 777–801. https://doi.org/10.1007/s13524-012-0193-x
REITSMA, M. B., FULLMAN, N., NG, M., SALAMA, J. S., ABAJOBIR, A., ABATE, K. H., ABBAFATI, C., ABERA, S. F., ABRAHAM, B. et al. (2017). Smoking prevalence and attributable disease burden in 195 countries and territories, 1990–2015: A systematic analysis from the global burden of disease study 2015. *Lancet* **389** 1885–1906.
ROSEN, L. (2013). An intuitive approach to understanding the attributable fraction of disease due to a risk factor: The case of smoking. *International Journal of Environmental Research and Public Health* **10** 2932–2943.
SHABANI, A., SEPASKHAH, A., KAMGAR-HAGHIGHI, A. and HONAR, T. (2018). Using double logistic equation to describe the growth of winter rapeseed. *The Journal of Agricultural Science* **156** 37–45.
SHIBUYA, K., INOUE, M. and LOPEZ, A. D. (2005). Statistical modeling and projections of lung cancer mortality in 4 industrialized countries. *Int. J. Cancer* **117** 476–485.
SIMONATO, L., AGUDO, A., AHRENS, W., BENHAMOU, E., BENHAMOU, S., BOFFETTA, P., BRENNAN, P., DARBY, S. C., FORASTIERE, F. et al. (2001). Lung cancer and cigarette smoking in Europe: An update of risk estimates and an assessment of inter-country heterogeneity. *Int. J. Cancer* **91** 876–887.
SMITH, T. R. and WAKEFIELD, J. (2016). A review and comparison of age-period-cohort models for cancer incidence. *Statist. Sci*. **31** 591–610. MR3598741 https://doi.org/10.1214/16-STS580
STOELDRAIJER, L., BONNEUX, L., VAN DUIN, C., VAN WISSEN, L. and JANSSEN, F. (2015). The future of smoking-attributable mortality: The case of England &Wales, Denmark and the Netherlands. *Addiction* **110** 336–345.
TACHFOUTI, N., RAHERISON, C., OBTEL, M. and NEJJARI, C. (2014). Mortality attributable to tobacco: Review of different methods. *Arch Public Health* **72** 22. https://doi.org/10.1186/2049-3258-72-22
TENG, A., ATKINSON, J., DISNEY, G., WILSON, N. and BLAKELY, T. (2017). Changing smoking-mortality association over time and across social groups: National census-mortality cohort studies from 1981 to 2011. *Sci. Rep*. **7** 11465. https://doi.org/10.1038/s41598-017-11785-x
UNITED NATIONS (2017). *World Population Prospects*. United Nations, New York, NY. Accessed: Oct. 15, 2018. Available at http://population.un.org/wpp/Download/Standard/Population/.

WANG, H. and PRESTON, S. H. (2009). Forecasting United States mortality using cohort smoking histories. *Proc. Natl. Acad. Sci. USA* **106** 393–398.

WEN, C., TSAI, S., CHEN, C., CHENG, T., TSAI, M. and LEVY, D. (2005). Smoking attributable mortality for Taiwan and its projection to 2020 under different smoking scenarios. *Tobacco Control* **14** i76–i80.

WOOD, S. N. (2003). Thin plate regression splines. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 95–114. MR1959095 https://doi.org/10.1111/1467-9868.00374

WORLD HEALTH ORGANIZATION (2017). Mortality Database. Last accessed: Oct. 15, 2018. Available at http://www.who.int/healthinfo/statistics/mortality_rawdata/en/.

YANG, X., MUSTARD, J. F., TANG, J. and XU, H. (2012). Regional-scale phenology modeling based on meteorological records and remote sensing observations. *Journal of Geophysical Research*: *Biogeosciences* **117**.

ZHENG, Y., JI, Y., DONG, H. and CHANG, C. (2018). The prevalence of smoking, second-hand smoke exposure, and knowledge of the health hazards of smoking among internal migrants in 12 provinces in China: A cross-sectional analysis. *BMC Public Health* **18** 655.

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 409–432
https://doi.org/10.1214/19-AOAS1311

MEASURING HUMAN ACTIVITY SPACES FROM GPS DATA WITH DENSITY RANKING AND SUMMARY CURVES

BY YEN-CHI CHEN* AND ADRIAN DOBRA†

*Department of Statistics, University of Washington, *yenchic@uw.edu; †adobra@uw.edu*

Activity spaces are fundamental to the assessment of individuals' dynamic exposure to social and environmental risk factors associated with multiple spatial contexts that are visited during activities of daily living. In this paper we survey existing approaches for measuring the geometry, size and structure of activity spaces, based on GPS data, and explain their limitations. We propose addressing these shortcomings through a nonparametric approach called density ranking and also through three summary curves: the mass-volume curve, the Betti number curve and the persistence curve. We introduce a novel mixture model for human activity spaces and study its asymptotic properties. We prove that the kernel density estimator, which at the present time, is one of the most widespread methods for measuring activity spaces, is not a stable estimator of their structure. We illustrate the practical value of our methods with a simulation study and with a recently collected GPS dataset that comprises the locations visited by 10 individuals over a six months period.

REFERENCES

APOSTOLOPOULOS, Y. and SONMEZ, S. (2007). *Population Mobility and Infectious Disease*. Springer, New York.

BASTA, L. A., RICHMOND, T. S. and WIEBE, D. J. (2010). Neighborhoods, daily activities, and measuring health risks experienced in urban environments. *Soc. Sci. Med.* **71** 1943–1950.

BERRIGAN, D., HIPP, J. A., HURVITZ, P. M., JAMES, P., JANKOWSKA, M. M., KERR, J., LADEN, F., LEONARD, T., MCKINNON, R. A. et al. (2015). Geospatial and contextual approaches to energy balance and health. *Ann. of GIS* **21** 157–168.

BERRY, E., CHEN, Y.-C., CISEWSKI-KEHE, J. and FASY, B. T. (2018). Functional summaries of persistence diagrams. Preprint. Available at arXiv:1804.01618.

BISCIO, C. A. N. and MØLLER, J. (2019). The accumulated persistence function, a new useful functional summary statistic for topological data analysis, with a view to brain artery trees and spatial point process applications. *J. Comput. Graph. Statist.* **28** 671–681. MR4007749 https://doi.org/10.1080/10618600.2019.1573686

BULIUNG, R. N. (2001). Spatiotemporal patterns of employment and non-work activities in Portland, Oregon. In *ESRI International User Conference*, San Diego, CA.

BULIUNG, R. N. and KANAROGLOU, P. S. (2006). Urban form and household activity-travel behavior. *Growth Change* **37** 172–199.

CADRE, B., PELLETIER, B. and PUDLO, P. (2013). Estimation of density level sets with a given probability content. *J. Nonparametr. Stat.* **25** 261–272. MR3039981 https://doi.org/10.1080/10485252.2012.750319

CARLSSON, G. (2009). Topology and data. *Bull. Amer. Math. Soc.* (*N.S.*) **46** 255–308. MR2476414 https://doi.org/10.1090/S0273-0979-09-01249-X

CHAIX, B., KESTENS, Y., PERCHOUX, C., KARUSISI, N., MERLO, J. and LABADI, K. (2012). An interactive mapping tool to assess individual mobility patterns in neighborhood studies. *Am. J. Prev. Med.* **43** 440–450.

CHAZAL, F. and MICHEL, B. (2017). An introduction to topological data analysis: Fundamental and practical aspects for data scientists. Preprint. Available at arXiv:1710.04019.

CHEN, Y.-C. (2019). Generalized cluster trees and singular measures. *Ann. Statist.* **47** 2174–2203. MR3953448 https://doi.org/10.1214/18-AOS1744

CHEN, Y.-C. and DOBRA, A. (2020). Supplement to "Measuring human activity spaces from GPS data with density ranking and summary curves." https://doi.org/10.1214/19-AOAS1311SUPP.

Key words and phrases. Activity space, global positioning systems (GPS), human mobility, kernel density estimation, space-time geography, topological data analysis.

CHEN, C., MA, J., SUSILO, Y., LIU, Y. and WANG, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transp. Res., Part C, Emerg. Technol.* **68** 285–299.
CHRISTIAN, W. J. (2012). Using geospatial technologies to explore activity-based retail food environments. *Spatial and Spatio-Temporal Epidemiology* **3** 287–295.
CLÉMENÇON, S. and JAKUBOWICZ, J. (2013). Scoring anomalies: A m-estimation formulation. In *Artificial Intelligence and Statistics* 659–667.
CLÉMENÇON, S. and THOMAS, A. (2018). Mass volume curves and anomaly ranking. *Electron. J. Stat.* **12** 2806–2872. MR3855357 https://doi.org/10.1214/18-EJS1474
CUMMINS, S., CURTIS, S., DIEZ-ROUX, A. V. and MACINTYRE, S. (2007). Understanding and representing 'place' in health research: A relational approach. *Soc. Sci. Med.* **65** 1825–1838.
DOBRA, A., WILLIAMS, N. E. and EAGLE, N. (2015). Spatiotemporal detection of unusual human population behavior using mobile phone data. *PLoS ONE* **10** 1–20.
DOBRA, A., BÄRNIGHAUSEN, T., VANDORMAEL, A. and TANSER, F. (2017). Space-time migration patterns and risk of HIV acquisition in rural South Africa. *AIDS* **31** 137–145.
DOSS, C. R. and WENG, G. (2018). Bandwidth selection for kernel density estimators of multivariate level sets and highest density regions. *Electron. J. Stat.* **12** 4313–4376. MR3892342 https://doi.org/10.1214/18-ejs1501
EDELSBRUNNER, H. and HARER, J. (2008). Persistent homology—a survey. In *Surveys on Discrete and Computational Geometry. Contemp. Math.* **453** 257–282. Amer. Math. Soc., Providence, RI. MR2405684 https://doi.org/10.1090/conm/453/08802
EDELSBRUNNER, H. and HARER, J. L. (2010). *Computational Topology*: *An Introduction*. Amer. Math. Soc., Providence, RI. MR2572029
ENTWISLE, B. (2007). Putting people into place. *Demography* **44** 687–703.
FAN, Y. and KHATTAK, A. (2008). Urban form, individual spatial footprints, and travel: Examination of space-use behavior. *Transp. Res. Rec.* **2082** 98–106.
GHRIST, R. (2008). Barcodes: The persistent topology of data. *Bull. Amer. Math. Soc.* (*N.S.*) **45** 61–75. MR2358377 https://doi.org/10.1090/S0273-0979-07-01191-3
GHRIST, R. (2014). *Elementary Applied Topology*. CreateSpace Independent Publishing Platform.
GOLLEDGE, R. G. (1999). Human wayfinding and cognitive maps. In *Wayfinding Behavior* (R. G. Golledge, ed.) 5–45. The Johns Hopkins Univ. Press, Baltimore, MD.
GOLLEDGE, R. G. and STIMSON, R. J. (1997). *Spatial Behavior*. The Guildford Press, New York.
HÄGERSTRAND, T. (1963). Geographic measurements of migration. In *Human Displacements*: *Measurement Methodological Aspects* (J. Sutter, ed.), Monaco.
HÄGERSTRAND, T. (1970). What about people in regional science? *Pap. Reg. Sci.* **24** 7–21.
HARDING, C., PATTERSON, Z. and MIRANDA-MORENO, L. F. (2013). Activity space geometry and its effect on mode choice. In *Transportation Research Board* 92*nd Annual Meeting*, Washington DC.
HARTIGAN, J. A. (1975). *Clustering Algorithms*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0405726
HIRSCH, J. A., WINTERS, M., CLARKE, P. and MCKAY, H. (2014). Generating GPS activity spaces that shed light upon the mobility habits of older adults: A descriptive analysis. *Int. J. Health Geogr.* **13** 51.
HURVITZ, P. M., MOUDON, A. V., KANG, B., SAELENS, B. E. and DUNCAN, G. E. (2014). Emerging technologies for assessing physical activity behaviors in space and time. *Front. Public Health* **2**.
KACZYNSKI, T., MISCHAIKOW, K. and MROZEK, M. (2004). *Computational Homology*. *Applied Mathematical Sciences* **157**. Springer, New York. MR2028588 https://doi.org/10.1007/b97315
KESTENS, Y., LEBEL, A., DANIEL, M., THÉRIAULT, M. and PAMPALON, R. (2010). Using experienced activity spaces to measure foodscape exposure. *Health Place* **16** 1094–1103.
KIM, S. and ULFARSSON, G. F. (2015). Activity space of older and working-age adults in the Puget Sound region. In *Transportation Research Board* 94*nd Annual Meeting*, Washington DC.
KWAN, M. P. (2000). Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: A methodological exploration with a large data set. *Transp. Res.* **8C** 185–203.
KWAN, M.-P. (2009). From place-based to people-based exposure measures. *Soc. Sci. Med.* **69** 1311–1313.
KWAN, M.-P. (2012). The uncertain geographic context problem. *Ann. Assoc. Am. Geogr.* **102** 958–968.
KWAN, M.-P. (2013). Beyond space (as we knew it): Toward temporally integrated geographies of segregation, health, and accessibility. *Ann. Assoc. Am. Geogr.* **103** 1078–1086.
LEE, J. H., DAVIS, A. W., YOON, S. Y. and GOULIAS, K. G. (2016). Activity space estimation with longitudinal observations of social media data. *Transp.* **43** 955–977.
LUM, P. Y., SINGH, G., LEHMAN, A., ISHKANOV, T., VEJDEMO-JOHANSSON, M., ALAGAPPAN, M., CARLSSON, J. and CARLSSON, G. (2013). Extracting insights from the shape of complex data using topology. *Sci. Rep.* **3** 1236.
MANAUGH, K. and EL-GENEIDY, A. M. (2012). What makes travel 'local': Defining and understanding local travel behavior. *J. Transp. Land Use* **5** 12–27.

MASON, D. M. and POLONIK, W. (2009). Asymptotic normality of plug-in level set estimates. *Ann. Appl. Probab.* **19** 1108–1142. MR2537201 https://doi.org/10.1214/08-AAP569

MATTHEWS, S. A. (2008). The salience of neighborhood. *Am. J. Prev. Med.* **34** 257–259.

MATTHEWS, S. A. (2011). Spatial polygamy and the heterogeneity of place: Studying people and place via egocentric methods. In *Communities, Neighborhoods, and Health: Expanding the Boundaries of Place* (L. M. Burton, S. A. Matthews, M. C. Leung, S. P. Kemp and D. T. Takeuchi, eds.) 35–55. Springer, New York, NY.

MATTHEWS, S. A. and YANG, T.-C. (2013). Spatial polygamy and contextual exposures (SPACEs): Promoting activity space approaches in research on place and health. *Am. Behav. Sci.* **57** 1057–1081.

MATTILA, P. (1995). *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability. Cambridge Studies in Advanced Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR1333890 https://doi.org/10.1017/CBO9780511623813

NEWSOME, T. H., WALCOTT, W. A. and SMITH, P. D. (1998). Urban activity spaces: Illustrations and application of a conceptual model for integrating the time and space dimensions. *Transp.* **25** 357–377.

NUÑEZ GARCIA, J., KUTALIK, Z., CHO, K.-H. and WOLKENHAUER, O. (2003). Level sets and minimum volume sets of probability density functions. *Internat. J. Approx. Reason.* **34** 25–47. MR2017778 https://doi.org/10.1016/S0888-613X(03)00052-5

PERCHOUX, C., CHAIX, B., CUMMINS, S. and KESTENS, Y. (2013). Conceptualization and measurement of environmental exposure in epidemiology: Accounting for activity space related to daily mobility. *Health Place* **21** 86–93.

POLONIK, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Process. Appl.* **69** 1–24. MR1464172 https://doi.org/10.1016/S0304-4149(97)00028-8

PREISS, D. (1987). Geometry of measures in $\mathbb{R}^n$: Distribution, rectifiability, and densities. *Ann. of Math.* (2) **125** 537–643. MR0890162 https://doi.org/10.2307/1971410

QIAO, W. (2017). Asymptotics and optimal bandwidth selection for nonparametric estimation of density level sets. Preprint. Available at arXiv:1707.09697.

RAI, R. K., BALMER, M., RIESER, M., VAZE, V. S., SCHÖNFELDER, S. and AXHAUSEN, K. W. (2007). Capturing human activity spaces: New geometries. *Transp. Res. Rec.* **2021** 70–80.

RICHARDSON, D. B., VOLKOW, N. D., KWAN, M.-P., KAPLAN, R. M., GOODCHILD, M. F. and CROYLE, R. T. (2013). Spatial turn in health research. *Science* **339** 1390–1392.

RIGOLLET, P. and VERT, R. (2009). Optimal rates for plug-in estimators of density level sets. *Bernoulli* **15** 1154–1178. MR2597587 https://doi.org/10.3150/09-BEJ184

RINALDO, A. and WASSERMAN, L. (2010). Generalized density clustering. *Ann. Statist.* **38** 2678–2722. MR2722453 https://doi.org/10.1214/10-AOS797

SCHÖNFELDER, S. and AXHAUSEN, K. W. (2003). Activity spaces: Measures of social exclusion? *Transp. Policy* **10** 273–286.

SCHÖNFELDER, S. and AXHAUSEN, K. W. (2004). On the variability of human activity spaces. In *The Real and Virtual Worlds of Spatial Planning* (M. Koll-Schretzenmayr, M. Keiner and G. Nussbaumer, eds.) 237–262. Springer, Berlin, Heidelberg.

SCOTT, C. D. and NOWAK, R. D. (2006). Learning minimum volume sets. *J. Mach. Learn. Res.* **7** 665–704. MR2274383

SHERMAN, J. E., SPENCER, J., PREISSER, J. S., GESLER, W. M. and ARCURY, T. A. (2005). A suite of methods for representing activity space in a healthcare accessibility study. *Int. J. Health Geogr.* **4** 24–24.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability*. CRC Press, London. MR0848134 https://doi.org/10.1007/978-1-4899-3324-9

WASSERMAN, L. (2016). Topological data analysis. Available at arXiv:1609.08227.

WASSERMAN, L. (2018). Topological data analysis. *Annu. Rev. Stat. Appl.* **5** 501–535. MR3774757 https://doi.org/10.1146/annurev-statistics-031017-100045

WIEHE, S. E., CARROLL, A. E., LIU, G. C., HABERKORN, K. L., HOCH, S. C., WILSON, J. S. and FORTENBERRY, J. D. (2008). Using gps-enabled cell phones to track the travel patterns of adolescents. *Int. J. Health Geogr.* **7** 22–22.

WIEHE, S., KWAN, M.-P., WILSON, J. and FORTENBERRY, J. (2013). Adolescent health-risk behavior and community disorder. *PLoS ONE* **8** e77667.

WILLIAMS, N. E., THOMAS, T. A., DUNBAR, M., EAGLE, N. and DOBRA, A. (2015). Measures of human mobility using mobile phone records enhanced with gis data. *PLoS ONE* **10** 1–16.

WONG, D. W. S. and SHAW, S.-L. (2011). Measuring segregation: An activity space approach. *J. Geogr. Syst.* **13** 127–145.

WORTON, B. J. (1987). A review of models of home range for animal movement. *Ecol. Model.* **38** 277–298.

ZENK, S. N., SCHULZ, A. J., MATTHEWS, S. A., ODOMS-YOUNG, A., WILBUR, J., WEGRZYN, L., GIBBS, K., BRAUNSCHWEIG, C. and STOKES, C. (2011). Activity space environment and dietary and physical activity behaviors: A pilot study. *Health Place* **17** 1150–1161.

ZENK, S. N., SCHULZ, A. J., ODOMS-YOUNG, A., WILBUR, J., MATTHEWS, S. A., GAMBOA, C., WEGRZYN, L. R., HOBSON, S. and STOKES, C. (2012). Feasibility of using global positioning systems (GPS) with diverse urban adults: Before and after data on perceived acceptability, barriers, and ease of use. *J. Phys. Act. Health* **9** 924–934.

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 433–451
https://doi.org/10.1214/19-AOAS1312

A COMPARISON OF PRINCIPAL COMPONENT METHODS BETWEEN MULTIPLE PHENOTYPE REGRESSION AND MULTIPLE SNP REGRESSION IN GENETIC ASSOCIATION STUDIES

BY ZHONGHUA LIU[1], IAN BARNETT[2] AND XIHONG LIN[3]

[1]*Department of Statistics and Actuarial Science, The University of Hong Kong, zhhliu@hku.hk*
[2]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, ibarnett@pennmedicine.upenn.edu*
[3]*Department of Biostatistics and Statistics, Harvard University, xlin@hsph.harvard.edu*

Principal component analysis (PCA) is a popular method for dimension reduction in unsupervised multivariate analysis. However, existing ad hoc uses of PCA in both multivariate regression (multiple outcomes) and multiple regression (multiple predictors) lack theoretical justification. The differences in the statistical properties of PCAs in these two regression settings are not well understood. In this paper we provide theoretical results on the power of PCA in genetic association testings in both multiple phenotype and SNP-set settings. The multiple phenotype setting refers to the case when one is interested in studying the association between a single SNP and multiple phenotypes as outcomes. The SNP-set setting refers to the case when one is interested in studying the association between multiple SNPs in a SNP set and a single phenotype as the outcome. We demonstrate analytically that the properties of the PC-based analysis in these two regression settings are substantially different. We show that the lower order PCs, that is, PCs with large eigenvalues, are generally preferred and lead to a higher power in the SNP-set setting, while the higher-order PCs, that is, PCs with small eigenvalues, are generally preferred in the multiple phenotype setting. We also investigate the power of three other popular statistical methods, the Wald test, the variance component test and the minimum p-value test, in both multiple phenotype and SNP-set settings. We use theoretical power, simulation studies, and two real data analyses to validate our findings.

REFERENCES

ASCHARD, H., VILHJÁLMSSON, B. J., GRELICHE, N., MORANGE, P. E., TRÉGOUËT, D. A. and KRAFT, P. (2014). Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am. J. Hum. Genet.* **94** 662–676.

CONNEELY, K. N. and BOEHNKE, M. (2007). So many correlated tests, so little time! Rapid adjustment of P-values for multiple correlated tests. *Am. J. Hum. Genet.* **81** 1158–1168.

HAN, B., KANG, H. M. and ESKIN, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet.* **5** e1000456. https://doi.org/10.1371/journal.pgen.1000456

HUANG, Y.-T. and LIN, X. (2013). Gene set analysis using variance component tests. *BMC Bioinform.* **14** 210.

HUNTER, D. J., KRAFT, P., JACOBS, K. B., COX, D. G., YEAGER, M., HANKINSON, S. E., WACHOLDER, S., WANG, Z., WELCH, R. et al. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39** 870–874.

JOLLIFFE, I. T. (1982). A note on the use of principal components in regression. *Appl. Stat.* 300–303.

KARACAÖREN, B., SILANDER, T., ÁLVAREZ-CASTRO, J. M., HALEY, C. S. and DE KONING, D. J. (2011). Association analyses of the MAS-QTL data set using grammar, principal components and Bayesian network methodologies. In *BMC Proceedings* **5** S8. BioMed Central Ltd.

KARASIK, D., CHEUNG, C. L., ZHOU, Y., CUPPLES, L. A., KIEL, D. P. and DEMISSIE, S. (2012). Genome-wide association of an integrated osteoporosis-related phenotype: Is there evidence for pleiotropic genes? *J. Bone Miner. Res.* **27** 319–330.

Key words and phrases. Dimension reduction, principal component analysis, eigen-values, hypothesis testing, multiple phenotypes, minimum p-value test, SNP-set, variance-component test.

LEE, S., ABECASIS, G. R., BOEHNKE, M. and LIN, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *Am. J. Hum. Genet.* **95** 5–23.

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

LIN, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika* **84** 309–326. MR1467049 https://doi.org/10.1093/biomet/84.2.309

LIU, Z. and LIN, X. (2018). Multiple phenotype association tests using summary statistics in genome-wide association studies. *Biometrics* **74** 165–175. MR3777937 https://doi.org/10.1111/biom.12735

LIU, Z. and LIN, X. (2019). A geometric perspective on the power of principal component association tests in multiple phenotype studies. *J. Amer. Statist. Assoc.* **114** 975–990. MR4011752 https://doi.org/10.1080/01621459.2018.1513363

MOSKVINA, V. and SCHMIDT, K. M. (2008). On multiple-testing correction in genome-wide association studies. *Genet. Epidemiol.* **32** 567–573.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81** 559–575.

SCHIFANO, E. D., LI, L., CHRISTIANI, D. C. and LIN, X. (2013). Genome-wide association analysis for multiple continuous secondary phenotypes. *Am. J. Hum. Genet.* **92** 744–759.

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.

STEPHENS, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PLoS ONE* **8** e65245. https://doi.org/10.1371/journal.pone.0065245

SUO, C., TOULOPOULOU, T., BRAMON, E., WALSHE, M., PICCHIONI, M., MURRAY, R. and OTT, J. (2013). Analysis of multiple phenotypes in genome-wide genetic mapping studies. *BMC Bioinform.* **14** 151. https://doi.org/10.1186/1471-2105-14-151

TESLOVICH, T. M., MUSUNURU, K., SMITH, A. V., EDMONDSON, A. C., STYLIANOU, I. M., KOSEKI, M., PIRRUCCELLO, J. P., RIPATTI, S., CHASMAN, D. I. et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466** 707–713.

WANG, K. and ABBOTT, D. (2008). A principal components regression approach to multilocus genetic association studies. *Genet. Epidemiol.* **32** 108–118.

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

ZHANG, F., GUO, X., WU, S., HAN, J., LIU, Y., SHEN, H. and DENG, H. W. (2012). Genome-wide pathway association studies of multiple correlated quantitative phenotypes using principle component analyses. *PLoS ONE* **7** e53320.

ZHOU, X. and STEPHENS, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies *Nat. Methods* **11** 407–409.

ZHU, X., FENG, T., TAYO, B. O., LIANG, J., YOUNG, J. H., FRANCESCHINI, N., SMITH, J. A., YANEK, L. R., SUN, Y. V. et al. (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet.* **96** 21–36.

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 452–472
https://doi.org/10.1214/19-AOAS1316

ESTIMATING CAUSAL EFFECTS IN STUDIES OF HUMAN BRAIN FUNCTION: NEW MODELS, METHODS AND ESTIMANDS

BY MICHAEL E. SOBEL[1] AND MARTIN A. LINDQUIST[2]

[1]*Department of Statistics, Columbia University, michael@stat.columbia.edu*
[2]*Department of Biostatistics, Johns Hopkins University, mlindqui@jhsph.edu*

Neuroscientists often use functional magnetic resonance imaging (fMRI) to infer effects of treatments on neural activity in brain regions. In a typical fMRI experiment, each subject is observed at several hundred time points. At each point, the blood oxygenation level dependent (BOLD) response is measured at 100,000 or more locations (voxels). Typically, these responses are modeled treating each voxel separately, and no rationale for interpreting associations as effects is given. Building on Sobel and Lindquist (*J. Amer. Statist. Assoc.* **109** (2014) 967–976), who used potential outcomes to define unit and average effects at each voxel and time point, we define and estimate both "point" and "cumulated" effects for brain regions. Second, we construct a multisubject, multivoxel, multirun whole brain causal model with explicit parameters for regions. We justify estimation using BOLD responses averaged over voxels within regions, making feasible estimation for all regions simultaneously, thereby also facilitating inferences about association between effects in different regions. We apply the model to a study of pain, finding effects in standard pain regions. We also observe more cerebellar activity than observed in previous studies using prevailing methods.

REFERENCES

BARBEY, A. K., KOENIGS, M. and GRAFMAN, J. (2013). Dorsolateral prefrontal contributions to human working memory. *Cortex* **49** 1195–1205.

BEAUCHAMP, M. S., LEE, K. E., HAXBY, J. V. and MARTIN, A. (2003). FMRI responses to video and point-light displays of moving humans and manipulable objects. *J. Cogn. Neurosci.* **15** 991–1001.

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

BORSOOK, D., MOULTON, E., TULLY, S., SCHMAHMANN, J. and BECERRA, L. (2008). Human cerebellar responses to brush and heat stimuli in healthy and neuropathic pain subjects. *Cerebellum* **7** 252–272.

BOWMAN, F. D. (2005). Spatio-temporal modeling of localized brain activity. *Biostatistics* **6** 558–575.

BOWMAN, F. D. (2007). Spatiotemporal models for region of interest analyses of functional neuroimaging data. *J. Amer. Statist. Assoc.* **102** 442–453. MR2370845 https://doi.org/10.1198/016214506000001347

BOWMAN, F. D., CAFFO, B., BASSETT, S. S. and KILTS, C. (2008). A Bayesian hierarchical framework for spatial modeling of fMRI data. *NeuroImage* **39** 146–156.

BOYNTON, G. M., ENGEL, S. A., GLOVER, G. H. and HEEGER, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *J. Neurosci.* **16** 4207–4221.

BUXTON, R. B. (2009). *Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques*. Cambridge Univ. Press, Cambridge.

CARP, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of FMRI experiments. *Front. Neurosci.* **6** 149. https://doi.org/10.3389/fnins.2012.00149

CORBETTA, M. and SHULMAN, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev., Neurosci.* **3** 201–215. https://doi.org/10.1038/nrn755

DAVIDIAN, M. and GILTINAN, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*, 1st ed. Chapman and Hall/CRC, New York.

DEGRAS, D. and LINDQUIST, M. A. (2014). A hierarchical model for simultaneous detection and estimation in multi-subject fMRI studies. *NeuroImage* **98** 61–72. https://doi.org/10.1016/j.neuroimage.2014.04.052

Key words and phrases. Causal inference, fMRI, functional connectivity, systematic error, pain, region of interest.

FRISTON, K. J. (2011). Functional and effective connectivity: A review. *Brain Connect*. **1** 13–36. https://doi.org/10.1089/brain.2011.0008

FRISTON, K. J., HARRISON, L. and PENNY, W. (2003). Dynamic causal modelling. *NeuroImage* **19** 1273–1302.

FRISTON, K. J., HOLMES, A. P., WORSLEY, K. J., POLINE, J.-P., FRITH, C. D. and FRACKOWIAK, R. S. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp*. **2** 189–210.

HARRISON, L. M. and GREEN, G. G. (2010). A Bayesian spatiotemporal model for very large data sets. *NeuroImage* **50** 1126–1141.

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc*. **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292

KONG, J., JENSEN, K., LOIOTILE, R., CHEETHAM, A., WEY, H.-Y., TAN, Y., ROSEN, B., SMOLLER, J. W., KAPTCHUK, T. J. et al. (2013). Functional connectivity of the frontoparietal network predicts cognitive modulation of pain. *Pain* **154** 459–467.

KWONG, K. K., BELLIVEAU, J. W., CHESLER, D. A., GOLDBERG, I. E., WEISSKOFF, R. M., PONCELET, B. P., KENNEDY, D. N., HOPPEL, B. E., COHEN, M. S. et al. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl. Acad. Sci. USA* **89** 5675–5679.

LINDQUIST, M. A. (2008). The statistical analysis of fMRI data. *Statist. Sci*. **23** 439–464. MR2530545 https://doi.org/10.1214/09-STS282

LINDQUIST, M. A. (2012). Functional causal mediation analysis with an application to brain connectivity. *J. Amer. Statist. Assoc*. **107** 1297–1309. MR3036396 https://doi.org/10.1080/01621459.2012.695640

LINDQUIST, M. A. and SOBEL, M. E. (2011). Graphical models, potential outcomes and causal inference: Comment on Ramsey, Spirtes and Glymour. *NeuroImage* **57** 334–336.

LINDQUIST, M. A. and SOBEL, M. E. (2013). Cloak and DAG: A response to the comments on our comment. *NeuroImage* **76** 446–449.

LINDQUIST, M. A. and SOBEL, M. E. (2016). Effective connectivity and causal inference in neuroimaging. In *Handbook of Neuroimaging Data Analysis* 419–440. CRC Press, Boca Raton.

LINDQUIST, M. A. and WAGER, T. D. (2007). Validity and power in hemodynamic response modeling: A comparison study and a new approach. *Hum. Brain Mapp*. **28** 764–784.

LINDQUIST, M. A., LOH, J. M., ATLAS, L. Y. and WAGER, T. D. (2009). Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. *NeuroImage* **45** S187–S198.

MCLNTOSH, A. and GONZALEZ-LIMA, F. (1994). Structural equation modeling and its application to network analysis in functional brain imaging. *Hum. Brain Mapp*. **2** 2–22.

MEJIA, A., YUE, Y. R., BOLIN, D., LINDREN, F. and LINDQUIST, M. A. (2017). A Bayesian general linear modeling approach to cortical surface fMRI data analysis. Preprint. Available at arXiv:1706.00959.

MIEZIN, F. M., MACCOTTA, L., OLLINGER, J., PETERSEN, S. and BUCKNER, R. (2000). Characterizing the hemodynamic response: Effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *NeuroImage* **11** 735–759.

MONTI, M. M. (2011). Statistical analysis of fMRI time-series: A critical review of the GLM approach. *Front. Human Neurosci*. **5** 28. https://doi.org/10.3389/fnhum.2011.00028

MOULTON, E. A., SCHMAHMANN, J. D., BECERRA, L. and BORSOOK, D. (2010). The cerebellum and pain: Passive integrator or active participator? *Brains Res. Rev.* **65** 14–27. https://doi.org/10.1016/j.brainresrev.2010.05.005

OGAWA, S., LEE, T.-M., KAY, A. R. and TANK, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci. USA* **87** 9868–9872.

PENNY, W. D., TRUJILLO-BARRETO, N. J. and FRISTON, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage* **24** 350–362.

POLDRACK, R. A. (2007). Region of interest analysis for fMRI. *Soc. Cogn. Affect. Neurosci*. **2** 67–70. https://doi.org/10.1093/scan/nsm006

POLDRACK, R. A., MUMFORD, J. A. and NICHOLS, T. E. (2011). *Handbook of Functional MRI Data Analysis*. Cambridge Univ. Press, Cambridge. MR2839490 https://doi.org/10.1017/CBO9780511895029

ROBINS, J. M. and HERNÁN, M. A. (2009). Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 553–599. CRC Press, Boca Raton, FL. MR1500133

ROEBROECK, A., FORMISANO, E. and GOEBEL, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *NeuroImage* **25** 230–242.

SANYAL, N. and FERREIRA, M. A. (2012). Bayesian hierarchical multi-subject multiscale analysis of functional MRI data. *NeuroImage* **63** 1519–1531.

SOBEL, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Amer. Statist. Assoc*. **101** 1398–1407. MR2307573 https://doi.org/10.1198/016214506000000636

SOBEL, M. E. and LINDQUIST, M. A. (2014). Causal inference for fMRI time series data with systematic errors of measurement in a balanced on/off study of social evaluative threat. *J. Amer. Statist. Assoc.* **109** 967–976. MR3265669 https://doi.org/10.1080/01621459.2014.922886

WAGER, T. D., VAZQUEZ, A., HERNANDEZ, L. and NOLL, D. C. (2005). Accounting for nonlinear BOLD effects in fMRI: Parameter estimates and a model for prediction in rapid event-related studies. *NeuroImage* **25** 206–218.

WAGER, T. D., ATLAS, L. Y., LINDQUIST, M. A., ROY, M., WOO, C.-W. and KROSS, E. (2013). An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368** 1388–1397.

WOO, C.-W., KRISHNAN, A. and WAGER, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage* **91** 412–419.

WOO, C.-W., ROY, M., BUHLE, J. T. and WAGER, T. D. (2015). Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. *PLoS Biol.* **13** e1002036. https://doi.org/10.1371/journal.pbio.1002036

WOOLRICH, M. W., JENKINSON, M., BRADY, J. M. and SMITH, S. M. (2004). Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE Trans. Med. Imag.* **23** 213–231.

YEO, B., KRIENEN, F. M., SEPULCRE, J., SABUNCU, M. R., LASHKARI, D., HOLLINSHEAD, M., ROFFMAN, J. L., SMOLLER, J. W., ZÖLLEI, L. et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106** 1125–1165.

ZHANG, L., GUINDANI, M., VERSACE, F., ENGELMANN, J. M. and VANNUCCI, M. (2016). A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data. *Ann. Appl. Stat.* **10** 638–666. MR3528355 https://doi.org/10.1214/16-AOAS926

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 473–493
https://doi.org/10.1214/19-AOAS1315

A HIERARCHICAL DEPENDENT DIRICHLET PROCESS PRIOR FOR MODELLING BIRD MIGRATION PATTERNS IN THE UK

BY ALEX DIANA[1,*], ELENI MATECHOU[1,†], JIM GRIFFIN[2] AND ALISON JOHNSTON[3]

[1]*School of Mathematics, Statistics and Actuarial Science, University of Kent,* [*]*ad603@kent.ac.uk;* [†]*e.matechou@kent.ac.uk*
[2]*Department of Statistical Science, University College London, j.griffin@ucl.ac.uk*
[3]*Cornell Lab of Ornithology, Cornell University, aj327@cornell.edu*

Environmental changes in recent years have been linked to phenological shifts which in turn are linked to the survival of species. The work in this paper is motivated by capture-recapture data on blackcaps collected by the British Trust for Ornithology as part of the Constant Effort Sites monitoring scheme. Blackcaps overwinter abroad and migrate to the UK annually for breeding purposes. We propose a novel Bayesian nonparametric approach for expressing the bivariate density of individual arrival and departure times at different sites across a number of years as a mixture model. The new model combines the ideas of the hierarchical and the dependent Dirichlet process, allowing the estimation of site-specific weights and year-specific mixture locations, which are modelled as functions of environmental covariates using a multivariate extension of the Gaussian process. The proposed modelling framework is extremely general and can be used in any context where multivariate density estimation is performed jointly across different groups and in the presence of a continuous covariate.

REFERENCES

ÁLVAREZ, M. A. and LAWRENCE, N. D. (2011). Computationally efficient convolved multiple output Gaussian processes. *J. Mach. Learn. Res.* **12** 1459–1500. MR2813145

BOTH, C., BOUWHUIS, S., LESSELLS, C. and VISSER, M. E. (2006). Climate change and population declines in a long-distance migratory bird. *Nature* **441** 81.

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, Z., WANG, B. and GORBAN, A. N. (2017). Multivariate Gaussian and Student-t process regression for multi-output prediction. ArXiv preprint. Available at arXiv:1703.04455.

DE IORIO, M., MÜLLER, P., ROSNER, G. L. and MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99** 205–215. MR2054299 https://doi.org/10.1198/016214504000000205

DIANA, A., MATECHOU, E., GRIFFIN, J. and JOHNSTON, A. (2020). Supplement to "A hierarchical dependent Dirichlet process prior for modelling bird migration patterns in the UK." https://doi.org/10.1214/19-AOAS1315SUPP.

DORAZIO, R. M., MUKHERJEE, B., ZHANG, L., GHOSH, M., JELKS, H. L. and JORDAN, F. (2008). Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics* **64** 635–644. MR2432438 https://doi.org/10.1111/j.1541-0420.2007.00873.x

EDDELBUETTEL, D., FRANÇOIS, R., ALLAIRE, J., USHEY, K., KOU, Q., RUSSEL, N., CHAMBERS, J. and BATES, D. (2011). Rcpp: Seamless R and C++ integration. *J. Stat. Softw.* **40** 1–18.

EGLINGTON, S. M., JULLIARD, R., GARGALLO, G., VAN DER JEUGD, H. P., PEARCE-HIGGINS, J. W., BAILLIE, S. R. and ROBINSON, R. A. (2015). Latitudinal gradients in the productivity of e uropean migrant warblers have not shifted northwards during a period of climate change. *Glob. Ecol. Biogeogr.* **24** 427–436.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

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

Key words and phrases. Hierarchical Dirichlet process, Gaussian process, capture recapture.

FORD, J. H., PATTERSON, T. A. and BRAVINGTON, M. V. (2015). Modelling latent individual heterogeneity in mark-recapture data with Dirichlet process priors. ArXiv preprint. Available at arXiv:1511.07103.

GE, H., CHEN, Y., WAN, M. and GHAHRAMANI, Z. (2015). Distributed inference for Dirichlet process mixture models. In *International Conference on Machine Learning* 2276–2284.

GIENAPP, P., LEIMU, R. and MERILÄ, J. (2007). Responses to climate change in avian migration time—microevolution versus phenotypic plasticity. *Clim. Res.* **35** 25–35.

GRIFFIN, J. E. and LEISEN, F. (2017). Compound random measures and their use in Bayesian non-parametrics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 525–545. MR3611758 https://doi.org/10.1111/rssb.12176

HJORT, N. L., HOLMES, C., MÜLLER, P. and WALKER, S. G. (2010). *Bayesian Nonparametrics* **28**. Cambridge Univ. Press, Cambridge.

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

JOHNSTON, A., ROBINSON, R. A., GARGALLO, G., JULLIARD, R., JEUGD, H. and BAILLIE, S. R. (2016). Survival of Afro-palaearctic passerine migrants in western Europe and the impacts of seasonal weather variables. *Ibis* **158** 465–480.

KINGMAN, J. F. C. (1967). Completely random measures. *Pacific J. Math.* **21** 59–78. MR0210185

KINGMAN, J. F. C. (1993). *Poisson Processes. Oxford Studies in Probability* **3**. The Clarendon Press, Oxford University Press, New York. MR1207584

MACEACHERN, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science* 50–5. Amer. Statist. Assoc., Alexandria, Va.

MACKENZIE, D. I., NICHOLS, J. D., LACHMAN, G. B., DROEGE, S., ROYLE, J. A. and LANGTIMM, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83** 2248–2255.

MANRIQUE-VALLIER, D. (2016). Bayesian population size estimation using Dirichlet process mixtures. *Biometrics* **72** 1246–1254. MR3591609 https://doi.org/10.1111/biom.12502

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

MØLLER, A. P., RUBOLINI, D. and LEHIKOINEN, E. (2008). Populations of migratory bird species that did not show a phenological response to climate change are declining. *Proc. Natl. Acad. Sci. USA*.

PEACH, W., BUCKLAND, S. and BAILLIE, S. (1996). The use of constant effort mist-netting to measure between-year changes in the abundance and productivity of common passerines. *Bird Study* **43** 142–156.

R CORE TEAM (2014). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RASMUSSEN, C. E. (2006). *Gaussian Processes in Machine Learning*. MIT Press.

ROBSON, D. and BARRIOCANAL, C. (2011). Ecological conditions in wintering and passage areas as determinants of timing of spring migration in trans-saharan migratory birds. *J. Anim. Ecol.* **80** 320–331.

ROYLE, J. A. (2004a). *N*-mixture models for estimating population size from spatially replicated counts. *Biometrics* **60** 108–115. MR2043625 https://doi.org/10.1111/j.0006-341X.2004.00142.x

ROYLE, J. A. (2004b). *N*-mixture models for estimating population size from spatially replicated counts. *Biometrics* **60** 108–115. MR2043625 https://doi.org/10.1111/j.0006-341X.2004.00142.x

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

TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. MR2279480 https://doi.org/10.1198/016214506000000302

WOLPERT, R. L. and ICKSTADT, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85** 251–267. MR1649114 https://doi.org/10.1093/biomet/85.2.251

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 494–517
https://doi.org/10.1214/19-AOAS1295

BAYESIAN MIXED EFFECTS MODELS FOR ZERO-INFLATED COMPOSITIONS IN MICROBIOME DATA ANALYSIS

BY BOYU REN[1,*], SERGIO BACALLADO[2], STEFANO FAVARO[3], TOMMI VATANEN[4], CURTIS HUTTENHOWER[1,4,†] AND LORENZO TRIPPA[1,‡]

[1]*Department of Biostatistics, Harvard University,* [*]*bor158@mail.harvard.edu;* [†]*chuttenh@hsph.harvard.edu;* [‡]*ltrippa@jimmy.harvard.edu*

[2]*Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, sb2116@cam.ac.uk*

[3]*Departimento di Scienze Economico-Sociali e Matematico-Statistiche, Università di Torino and Collegio Carlo Alberto, stefano.favaro@unito.it*

[4]*Liggins Institute, University of Auckland, t.vatanen@auckland.ac.nz*

Detecting associations between microbial compositions and sample characteristics is one of the most important tasks in microbiome studies. Most of the existing methods apply univariate models to single microbial species separately, with adjustments for multiple hypothesis testing. We propose a Bayesian analysis for a generalized mixed effects linear model tailored to this application. The marginal prior on each microbial composition is a Dirichlet process, and dependence across compositions is induced through a linear combination of individual covariates, such as disease biomarkers or the subject's age, and latent factors. The latent factors capture residual variability and their dimensionality is learned from the data in a fully Bayesian procedure. The proposed model is tested in data analyses and simulation studies with zero-inflated compositions. In these settings and within each sample, a large proportion of counts per microbial species are equal to zero. In our Bayesian model a priori the probability of compositions with absent microbial species is strictly positive. We propose an efficient algorithm to sample from the posterior and visualizations of model parameters which reveal associations between covariates and microbial compositions. We evaluate the proposed method in simulation studies, and then analyze a microbiome dataset for infants with type 1 diabetes which contains a large proportion of zeros in the sample-specific microbial compositions.

REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* **11** R106. https://doi.org/10.1186/gb-2010-11-10-r106

ARBEL, J., MENGERSEN, K. and ROUSSEAU, J. (2016). Bayesian nonparametric dependent model for partially replicated data: The influence of fuel spills on species diversity. *Ann. Appl. Stat.* **10** 1496–1516. MR3553233 https://doi.org/10.1214/16-AOAS944

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. MR2806429 https://doi.org/10.1093/biomet/asr013

BORG, I. and GROENEN, P. J. F. (2005). *Modern Multidimensional Scaling*: *Theory and Applications*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2158691

BROOKS, S. P. and GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7** 434–455. MR1665662 https://doi.org/10.2307/1390675

CHEN, J. and LI, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* **7** 418–442. MR3086425 https://doi.org/10.1214/12-AOAS592

FANARO, S., CHIERICI, R., GUERRINI, P. and VIGI, V. (2003). Intestinal microflora in early infancy: Composition and development. *Acta Pdœiatr.* **92** 48–55.

Key words and phrases. Truncated dependent Dirichlet processes, latent factor model, type 1 diabetes.

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

GEVERS, D., KUGATHASAN, S., DENSON, L. A., VÁZQUEZ-BAEZA, Y., VAN TREUREN, W., REN, B., SCHWAGER, E., KNIGHTS, D., SONG, S. J. et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host & Microbe* **15** 382–392. https://doi.org/10.1016/j.chom.2014.02.005

GRANTHAM, N. S., GUAN, Y., REICH, B. J., BORER, E. T. and GROSS, K. (2019). MIMIX: A Bayesian mixed-effects model for microbiome data from designed experiments. *J. Amer. Statist. Assoc.* **0** 1–16. https://doi.org/10.1080/01621459.2019.1626242

GREENBLUM, S., TURNBAUGH, P. J. and BORENSTEIN, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc. Natl. Acad. Sci. USA* **109** 594–599.

GRIFFIN, J. E., KOLOSSIATIS, M. and STEEL, M. F. J. (2013). Comparing distributions by using dependent normalized random-measure mixtures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 499–529. MR3065477 https://doi.org/10.1111/rssb.12002

HUMAN MICROBIOME PROJECT CONSORTIUM (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486** 207–214.

ISHWARAN, H. and ZAREPOUR, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canad. J. Statist.* **30** 269–283. MR1926065 https://doi.org/10.2307/3315951

JAMES, L. F., LIJOI, A. and PRÜNSTER, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.* **36** 76–97. MR2508332 https://doi.org/10.1111/j.1467-9469.2008.00609.x

JOHNSON, D. S., REAM, R. R., TOWELL, R. G., WILLIAMS, M. T. and LEON GUERRERO, J. D. (2013). Bayesian clustering of animal abundance trends for inference and dimension reduction. *J. Agric. Biol. Environ. Stat.* **18** 299–313. MR3110895 https://doi.org/10.1007/s13253-013-0143-0

KOHAVI, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the* 14*th International Joint Conference on Artificial Intelligence*, *Vol.* 2. *IJCAI'*95 1137–1143. Morgan Kaufmann Publishers Inc., San Francisco, CA.

KOSTIC, A. D., GEVERS, D., SILJANDER, H., VATANEN, T., HYÖTYLÄINEN, T., HÄMÄLÄINEN, A.-M., PEET, A., TILLMANN, V., PÖHÖ, P. et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host & Microbe* **17** 260–273. https://doi.org/10.1016/j.chom.2015.01.001

LEDOUX, M. and TALAGRAND, M. (2011). *Probability in Banach Spaces*: *Isoperimetry and Processes*. *Classics in Mathematics*. Springer, Berlin. Reprint of the 1991 edition. MR2814399

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

LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20** 1260–1291. MR3217444 https://doi.org/10.3150/13-BEJ521

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34** 1–41. MR0415861

LOZUPONE, C., LLADSER, M. E., KNIGHTS, D., STOMBAUGH, J. and KNIGHT, R. (2011). UniFrac: An effective distance metric for microbial community comparison. *ISME J.* **5** 169–172.

MACEACHERN, S. N. (2000). Dependent Dirichlet processes. Technical Report, Dept. Statistics, The Ohio State Univ.

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

MÜLLER, P., QUINTANA, F. and ROSNER, G. L. (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist.* **20** 260–278. Supplementary material available online. MR2816548 https://doi.org/10.1198/jcgs.2011.09066

PAULSON, J. N., STINE, O. C., BRAVO, H. C. and POP, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10** 1200–1202.

QIN, J., LI, R., RAES, J., ARUMUGAM, M., BURGDORF, K. S., MANICHANH, C., NIELSEN, T., PONS, N., LEVENEZ, F. et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464** 59–65.

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

REN, B., BACALLADO, S., FAVARO, S., HOLMES, S. and TRIPPA, L. (2017). Bayesian nonparametric ordination for the analysis of microbial communities. *J. Amer. Statist. Assoc.* **112** 1430–1442. MR3750866 https://doi.org/10.1080/01621459.2017.1288631

REN, B., BACALLADO, S., FAVARO, S., VATANEN, T., HUTTENHOWER, C. and TRIPPA, L. (2020). Supplement to "Bayesian mixed effects models for zero-inflated compositions in microbiome data analysis." https://doi.org/10.1214/19-AOAS1295SUPPA, https://doi.org/10.1214/19-AOAS1295SUPPB.

ROBERT, P. and ESCOUFIER, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **25** 257–265. MR0440801 https://doi.org/10.2307/2347233

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

RODRÍGUEZ, A. and DUNSON, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal.* **6** 145–177. MR2781811 https://doi.org/10.1214/11-BA605

SEGATA, N., BÖRNIGEN, D., MORGAN, X. C. and HUTTENHOWER, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4** 2304. https://doi.org/10.1038/ncomms3304

TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101** 1566–1581. https://doi.org/10.1198/016214506000000302

VATANEN, T., KOSTIC, A. D., D'HENNEZEL, E., SILJANDER, H., FRANZOSA, E. A., YASSOUR, M., KOLDE, R., VLAMAKIS, H., ARTHUR, T. D. et al. (2016). Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165** 842–853. https://doi.org/10.1016/j.cell.2016.04.007

VEHTARI, A., GELMAN, A. and GABRY, J. (2015). Pareto smoothed importance sampling. arXiv preprint arXiv:1507.02646.

VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27** 1413–1432. MR3647105 https://doi.org/10.1007/s11222-016-9696-4

WADSWORTH, W. D., ARGIENTO, R., GUINDANI, M., GALLOWAY-PENA, J., SHELBURNE, S. A. and VANNUCCI, M. (2017). An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinform.* **18** 94. https://doi.org/10.1186/s12859-017-1516-0

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

XU, L., PATERSON, A. D., TURPIN, W. and XU, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PLoS ONE* **10** 1–30. https://doi.org/10.1371/journal.pone.0129606

The Annals of Applied Statistics
2020, Vol. 14, No. 1, 518–520
https://doi.org/10.1214/19-AOAS1320
Main article: https://doi.org/10.1214/16-AOAS1001

CORRECTION: SENSITIVITY ANALYSIS FOR AN UNOBSERVED MODERATOR IN RCT-TO-TARGET-POPULATION GENERALIZATION OF TREATMENT EFFECTS

BY TRANG QUYNH NGUYEN[1] AND ELIZABETH A. STUART[2]

[1]*Department of Mental Health, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, trang.nguyen@jhu.edu*

[2]*Department of Mental Health, Department of Biostatistics, Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, estuart@jhu.edu*

REFERENCES

NGUYEN, T. Q., EBNESAJJAD, C., COLE, S. R. and STUART, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Ann. Appl. Stat.* **11** 225–247. MR3634322 https://doi.org/10.1214/16-AOAS1001

NGUYEN, T. Q., ACKERMAN, B., SCHMID, I., COLE, S. R. and STUART, E. A. (2018). Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *PLoS ONE* **13** e0208795. https://doi.org/10.1371/journal.pone.0208795

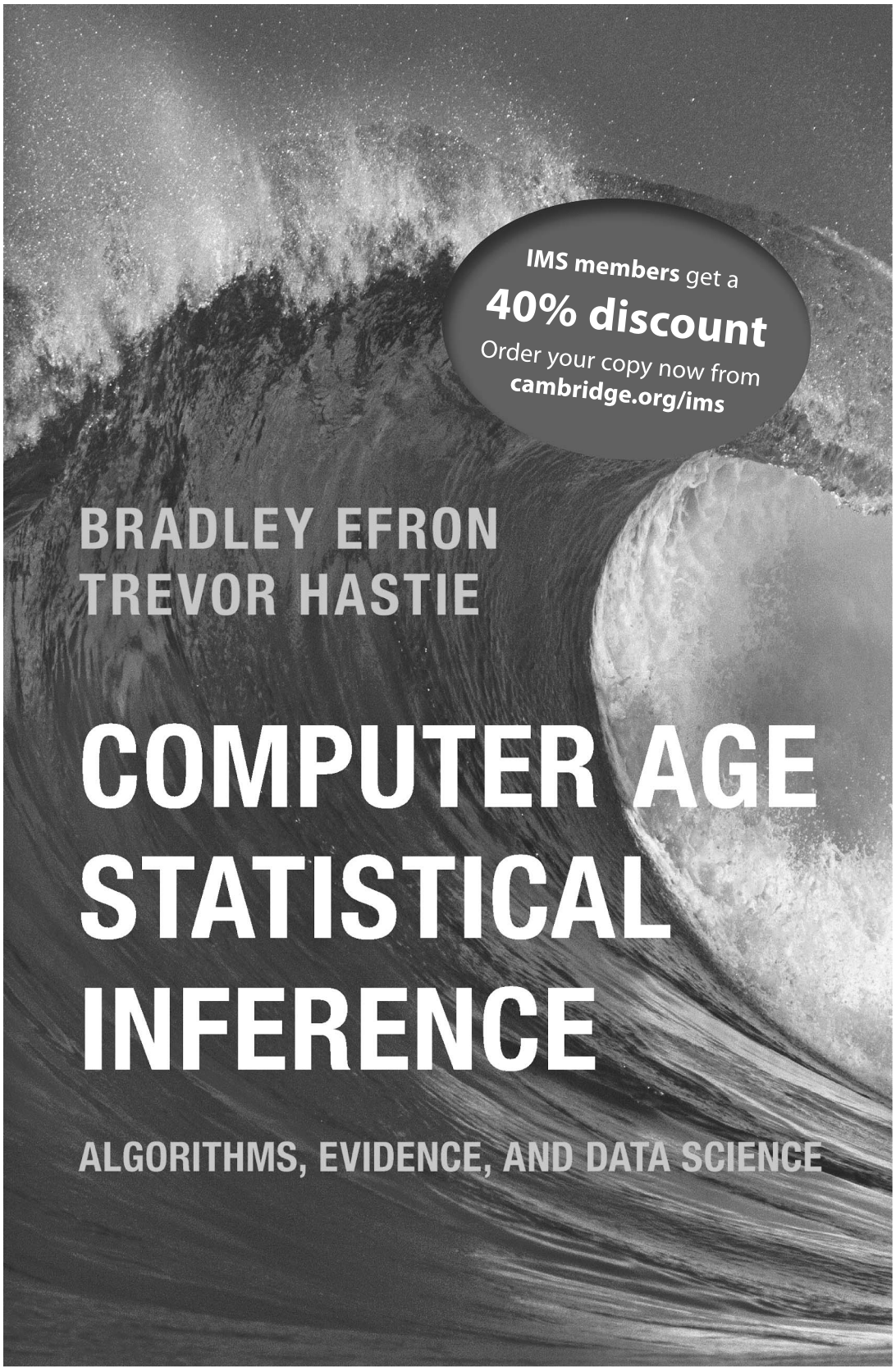
The Annals of Applied Statistics

Next Issues

Continued

The Annals of Applied Statistics

Next Issues–*Continued*


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