

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

- Climate inference on daily rainfall across the Australian continent, 1876–2015
MICHAEL BERTOLACCI, EDWARD CRIPPS, ORI ROSEN, JOHN W. LAU AND SALLY CRIPPS 683
- Modelling ocean temperatures from bio-probes under preferential sampling
DANIEL DINSDALE AND MATIAS SALIBIAN-BARRERA 713
- Complete spatial model calibration YEN-NING HUANG, BRIAN J. REICH, MONTSERRAT FUENTES AND A. SANKARASUBRAMANIAN 746
- Extended sensitivity analysis for heterogeneous unmeasured confounding with an application to sibling studies of returns to education
COLIN B. FOGARTY AND RAIDEN B. HASEGAWA 767
- A hierarchical multivariate spatio-temporal model for clustered climate data with annual cycles GIANLUCA MASTRANTONIO, GIOVANNA JONA LASINIO, ALESSIO POLLICE, GIULIA CAPOTORTI, LORENZO TEODONIO, GIULIO GENOVA AND CARLO BLASI 797
- Fused comparative intervention scoring for heterogeneity of longitudinal intervention effects JARED D. HULING, MENGANG YU AND MAUREEN SMITH 824
- Graphical models for zero-inflated single cell gene expression ANDREW MCDAVID, RAPHAEL GOTTARDO, NOAH SIMON AND MATHIAS DRTON 848
- TreeClone: Reconstruction of tumor subclone phylogeny based on mutation pairs using next generation sequencing data TIANJIAN ZHOU, SUBHAJIT SENGUPTA, PETER MÜLLER AND YUAN JI 874
- Latent space modelling of multidimensional networks with application to the exchange of votes in Eurovision song contest SILVIA D'ANGELO, THOMAS BRENDAN MURPHY AND MARCO ALFÒ 900
- Modeling association in microbial communities with clique loglinear models
ADRIAN DOBRA, CAMILO VALDES, DRAGANA AJDIC, BERTRAND CLARKE AND JENNIFER CLARKE 931
- Variable prioritization in nonlinear black box methods: A genetic association case study
LORIN CRAWFORD, SETH R. FLAXMAN, DANIEL E. RUNCIE AND MIKE WEST 958
- Coherence-based time series clustering for statistical inference and visualization of brain connectivity CAROLINA EUÁN, YING SUN AND HERNANDO OMBAO 990
- Sparse principal component analysis with missing observations
SEYOUNG PARK AND HONGYU ZHAO 1016
- Adaptive gPCA: A method for structured dimensionality reduction with applications to microbiome data JULIA FUKUYAMA 1043
- Learning algorithms to evaluate forensic glass evidence
SEYOUNG PARK AND ALICIA CARRIQUIRY 1068
- Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition MIAOYAN WANG, JONATHAN FISCHER AND YUN S. SONG 1103

continued

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—Continued from front cover

- Nonparametric testing for differences in electricity prices: The case of the Fukushima nuclear accident DOMINIK LIEBL 1128
- Nonparametric inference for immune response thresholds of risk in vaccine studies
KEVIN M. DONOVAN, MICHAEL G. HUDGENS AND PETER B. GILBERT 1147
- Semiparametric empirical best prediction for small area estimation of unemployment indicators MARIA FRANCESCA MARINO, MARIA GIOVANNA RANALLI, NICOLA SALVATI AND MARCO ALFÒ 1166
- The identity of the zero-truncated, one-inflated likelihood and the zero-one-truncated likelihood for general count densities with an application to drink-driving in Britain
DANKMAR BÖHNING AND PETER G. M. VAN DER HEIJDEN 1198
- Phylogeny-based tumor subclone identification using a Bayesian feature allocation model
LI ZENG, JOSHUA L. WARREN AND HONGYU ZHAO 1212
- Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality
RACHEL C. NETHERY, FABRIZIA MEALLI AND FRANCESCA DOMINICI 1242
- Survival analysis of DNA mutation motifs with penalized proportional hazards
JEAN FENG, DAVID A. SHAW, VLADIMIR N. MININ, NOAH SIMON AND
FREDERICK A. MATSEN IV 1268
- Early diagnosis of neurological disease using peak degeneration ages of multiple biomarkers FEI GAO, YUANJIA WANG, DONGLIN ZENG AND
FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE 1295
- Correction to: Statistical modeling and analysis of trace element concentrations in forensic glass evidence KAREN D. H. PAN AND KAREN KAFADAR 1319

THE ANNALS OF APPLIED STATISTICS

Vol. 13, No. 2, pp. 683–1349 June 2019

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

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

Past President: Alison Etheridge, Department of Statistics, University of Oxford, Oxford, OX1 3LB, United Kingdom

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 13, Number 2, June 2019. 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.

CLIMATE INFERENCE ON DAILY RAINFALL ACROSS THE AUSTRALIAN CONTINENT, 1876–2015

BY MICHAEL BERTOLACCI^{*,1}, EDWARD CRIPPS^{*,1}, ORI ROSEN^{‡,2},
 JOHN W. LAU^{*} AND SALLY CRIPPS^{†,3,4}

University of Western Australia^{}, University of Sydney[†] and
 University of Texas at El Paso[‡]*

Daily precipitation has an enormous impact on human activity, and the study of how it varies over time and space, and what global indicators influence it, is of paramount importance to Australian agriculture. We analyze over 294 million daily rainfall measurements since 1876, spanning 17,606 sites across continental Australia. The data are not only large but also complex, and the topic would benefit from a common and publicly available statistical framework. We propose a Bayesian hierarchical mixture model that accommodates mixed discrete-continuous data. The observational level describes site-specific temporal and climatic variation via a mixture-of-experts model. At the next level of the hierarchy, spatial variability of the mixture weights' parameters is modeled by a spatial Gaussian process prior. A parallel and distributed Markov chain Monte Carlo sampler is developed which scales the model to large data sets. We present examples of posterior inference on the mixture weights, monthly intensity levels, daily temporal dependence, offsite prediction of the effects of climate drivers and long-term rainfall trends across the entire continent. Computer code implementing the methods proposed in this paper is available as an R package.

REFERENCES

- AMDAHL, G. M. (1967). Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18–20, 1967, Spring Joint Computer Conference. AFIPS '67 (Spring)* 483–485. ACM, New York.
- ANDRICH, M. A. and IMBERGER, J. (2013). The effect of land clearing on rainfall and fresh water resources in western Australia: A multi-functional sustainability analysis. *J. Appl. Econometrics* **20** 549–563.
- BERTOLACCI, M., CRIPPS, E., ROSEN, O., LAU, J. and CRIPPS, S. (2019a). Conditional distributions for the sampling scheme in “Climate inference on daily rainfall across the Australian continent, 1876–2015.” DOI:[10.1214/18-AOAS1218SUPPA](https://doi.org/10.1214/18-AOAS1218SUPPA).
- BERTOLACCI, M., CRIPPS, E., ROSEN, O., LAU, J. and CRIPPS, S. (2019b). Model comparison supplement for “Climate inference on daily rainfall across the Australian continent, 1876–2015.” DOI:[10.1214/18-AOAS1218SUPPB](https://doi.org/10.1214/18-AOAS1218SUPPB).
- BERTOLACCI, M., CRIPPS, E., ROSEN, O., LAU, J. and CRIPPS, S. (2019c). Model diagnostics for “Climate inference on daily rainfall across the Australian continent, 1876–2015.” DOI:[10.1214/18-AOAS1218SUPPC](https://doi.org/10.1214/18-AOAS1218SUPPC).

Key words and phrases. Climate, rainfall, Australia, mixture-of-experts, Gaussian processes, parallel and distributed computing.

- CAI, W., VAN RENSCH, P., COWAN, T. and HENDON, H. (2012). An asymmetry in the IOD and ENSO teleconnection pathway and its impact on Australian climate. *J. Climate* **25** 6318–6329.
- CHARLES, S. P., BATES, B. C. and HUGHES, J. P. (1999). A spatiotemporal model for downscaling precipitation occurrence and amounts. *J. Geophys. Res., Atmos.* **104** 31657–31669.
- CHIB, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *J. Econometrics* **75** 79–97. [MR1414504](#)
- COMPO, G. P., WHITAKER, J. S., SARDESHMUKH, P. D., MATSUI, N., ALLAN, R. J., YIN, X., GLEASON, B. E., VOSE, R. S., RUTLEDGE, G., BESSEMOULIN, P. et al. (2011). The twentieth century reanalysis project. *Q. J. R. Meteorol. Soc.* **137** 1–28.
- DAMSLETH, E. (1975). Conjugate classes for gamma distributions. *Scand. J. Stat.* **2** 80–84. [MR0378169](#)
- EDDELBUEITTEL, D. and SANDERSON, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Comput. Statist. Data Anal.* **71** 1054–1063. [MR3132026](#)
- FENG, J., LI, J. and LI, Y. (2010). Is there a relationship between SAM and southwest western Australia winter rainfall. *J. Climate* **23** 6082–6089.
- FURRER, E. M. and KATZ, R. W. (2007). Generalized linear modeling approach to stochastic weather generators. *Climate Research* **34** 129–144.
- GONG, D. and WANG, S. (1999). Definition of Antarctic oscillation index. *Geophysical Research Letters* **26** 459–462.
- HENDON, H., THOMPSON, D. and WHEELER, M. (2007). Australian rainfall and surface temperature variations associated with the southern hemisphere annular mode. *J. Climate* **20** 2452–2467.
- HOLSCLAW, T., GREENE, A. M., ROBERTSON, A. W. and SMYTH, P. (2016). A Bayesian hidden Markov model of daily precipitation over South and East Asia. *Journal of Hydrometeorology* **17** 3–25.
- HOLSCLAW, T., GREENE, A. M., ROBERTSON, A. W. and SMYTH, P. (2017). Bayesian nonhomogeneous Markov models via Pólya-gamma data augmentation with applications to rainfall modeling. *Ann. Appl. Stat.* **11** 393–426. [MR3634329](#)
- JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. and HINTON, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput.* **3** 79–87.
- KALA, J., LYONS, T. J. and NAIR, U. S. (2011). Numerical simulations of the impacts of land-cover change on cold fronts in South-West western Australia. *Boundary-Layer Meteorology* **138** 121–138.
- KING, A., ALEXANDER, L. and DONAT, M. (2013). Asymmetry in the response of eastern Australia extreme rainfall to low-frequency Pacific variability. *Geophysical Research Letters* **40** 1–7.
- KLEIBER, W., KATZ, R. W. and RAJAGOPALAN, B. (2012). Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water Resour. Res.* **48**.
- LYNCH, N. A. (1996). *Distributed Algorithms. The Morgan Kaufmann Series in Data Management Systems*. Morgan Kaufmann, San Francisco, CA. [MR1388778](#)
- NAVEAU, P., HUSER, R., RIBEREAU, P. and HANNART, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resour. Res.* **52** 2753–2769.
- PEPLER, A., TIMBAL, B., RAKICH, C. and COUTTS-SMITH, A. (2014). Indian Ocean dipole overrides ENSO’s influence on cool season rainfall across the eastern seaboard of Australia. *J. Climate* **27** 3816–3826.
- PITMAN, A. J., NARISMA, G. G., PIELKE, R. A. and HOLBROOK, N. J. (2004). The impact of land cover change on the climate of southwest western Australia. *J. Geophys. Res.* **109** 1–12.
- POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. [MR3174712](#)
- R CORE TEAM (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

- RAYNER, N. A., PARKER, D. E., HORTON, E. B., FOLLAND, C. K., ALEXANDER, L. V., ROWELL, D. P., KENT, E. C. and KAPLAN, A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res., Atmos.* **108**.
- RICHARDSON, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resour. Res.* **17** 182–190.
- RISBEY, J., POOK, M., MCINTOSH, P., WHEELER, M. and HENDON, H. (2009). On the remote drivers of rainfall variability in Australia. *Mon. Weather Rev.* **137** 3233–3253.
- ROSEN, O., STOFFER, D. S. and WOOD, S. (2009). Local spectral analysis via a Bayesian mixture of smoothing splines. *J. Amer. Statist. Assoc.* **104** 249–262. [MR2504376](#)
- SAJI, N. H., GOSWAMI, B. N., VINAYACHANDRAN, P. N. and YAMAGATA, T. (1999). A dipole mode in the tropical Indian Ocean. *Nature* **401** 360–363.
- STERN, R. D. and COE, R. (1984). A model fitting analysis of daily rainfall data. *J. R. Stat. Soc., A* **147** 1–34.
- STONE, R. (2014). Constructing a framework for national drought policy: The way forward—The way Australia developed and implemented the national drought policy. *Weather and Climate Extremes* **3** 117–125.
- TROUP, A. J. (1965). The “southern oscillation”. *Q. J. R. Meteorol. Soc.* **91** 490–506.
- UMMENHOFER, C. C., ENGLAND, M. H., MCINTOSH, P. C., MEYERS, G. A., POOK, M. J., RISBEY, J. S., GUPTA, A. S. and TASCHETTO, A. S. (2009). What causes southeast Australia’s worst droughts?. *Geophysical Research Letters* **36** L04707.
- UMMENHOFER, C. C., GUPTA, A. S., BRIGGS, P. R., ENGLAND, M. H., MCINTOSH, P. C., MEYERS, G. A., POOK, M. J., RAUPACH, M. R. and RISBEY, J. S. (2011). Indian and Pacific Ocean influences on southeast Australian drought and soil moisture. *J. Climate* **24** 1313–1336.
- UMMENHOFER, C. C., GUPTA, A. S., ENGLAND, M. H., TASCHETTO, A. S., BRIGGS, P. R. and RAUPACH, M. R. (2015). How did ocean warming affect Australian rainfall extremes during the 2010/2011 La Niña event. *Geophysical Letters* **42** 9942–9951.
- VAN DIJK, A., BECK, H., CROSBIE, R., DE JEU, R., LIU, G., PODGER, Y., TIMBAL, B. and VINEY, N. (2013). The Millennium Drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society. *Water Resour. Res.* **49**.
- VRAC, M. and NAVEAU, P. (2007). Stochastic downscaling of precipitation: From dry events to heavy rainfalls. *Water Resour. Res.* **43**.
- WAHBA, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics **59**. SIAM, Philadelphia, PA. [MR1045442](#)
- WILKS, D. S. (1999). Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and Forest Meteorology* **93** 153–169.
- WOOD, S. (2013). Applications of Bayesian smoothing splines. In *Bayesian Theory and Applications* (P. Damien, P. Dellaportas, N. G. Polson and D. A. Stephens, eds.) 309–335. Oxford Univ. Press, Oxford. [MR3221170](#)
- WOOD, S., ROSEN, O. and KOHN, R. (2011). Bayesian mixtures of autoregressive models. *J. Comput. Graph. Statist.* **20** 174–195. [MR2816544](#)
- YU, H. (2002). Rmpi: Parallel statistical computing in R. *R News* **2** 10–14.

MODELLING OCEAN TEMPERATURES FROM BIO-PROBES UNDER PREFERENTIAL SAMPLING

BY DANIEL DINSDALE AND MATIAS SALIBIAN-BARRERA

University of British Columbia

In the last 25 years there has been an important increase in the amount of data collected from animal-mounted sensors (bio-probes) which are often used to study the animals' behaviour or environment. We focus here on an example of the latter, where the interest is in sea surface temperature (SST), and measurements are taken from sensors mounted on elephant seals in the southern Indian Ocean. We show that standard geostatistical models may not be reliable for this type of data, due to the possibility that the regions visited by the animals may depend on the SST. This phenomenon is known in the literature as preferential sampling, and, if ignored, it may affect the resulting spatial predictions and parameter estimates. Research on this topic has been mostly restricted to stationary sampling locations such as monitoring sites. The main contribution of this manuscript is to extend this methodology to observations obtained by devices that move through the region of interest, as is the case with the tagged seals. More specifically, we propose a flexible framework for inference on preferentially sampled fields where the process that generates the sampling locations is stochastic and moving over time through a two-dimensional space. Our simulation studies confirm that predictions obtained from the preferential sampling model are more reliable when this phenomenon is present, and they compare very well to the standard ones when there is no preferential sampling. Finally, we note that the conclusions of our analysis of the SST data can change considerably when we incorporate preferential sampling in the model.

REFERENCES

- ALBERTSEN, C. M., WHORISKEY, K., YURKOWSKI, D., NIELSEN, A. and FLEMMING, J. M. (2015). Fast fitting of non-Gaussian state-space models to animal movement data via template model builder. *Ecology* **96** 2598–2604.
- AUGER-MÉTHÉ, M., FIELD, C., ALBERTSEN, C. M., DEROCHER, A. E., LEWIS, M. A., JONSEN, I. D. and FLEMMING, J. M. (2016). State-space models' dirty little secrets: Even simple linear Gaussian models can have estimation problems. *Sci. Rep.* **6** 26677.
- AUGER-MÉTHÉ, M., ALBERTSEN, C. M., JONSEN, I. D., DEROCHER, A. E., LIDGARD, D. C., STUDHOLME, K. R., BOWEN, W. D., CROSSIN, G. T. and FLEMMING, J. M. (2017). Spatiotemporal modelling of marine movement data using Template Model Builder (TMB). *Mar. Ecol. Prog. Ser.* **565** 237–249.
- BANERJEE, S. and GELFAND, A. E. (2003). On smoothness properties of spatial processes. *J. Multivariate Anal.* **84** 85–100. [MR1965824](#)

Key words and phrases. Animal movement models, bio-probes, Laplace approximation, preferential sampling, template model builder.

- BANERJEE, S., GELFAND, A. E. and SIRMANS, C. F. (2003). Directional rates of change under spatial process models. *J. Amer. Statist. Assoc.* **98** 946–954. [MR2041483](#)
- BOLKER, B. M., GARDNER, B., MAUNDER, M., BERG, C. W., BROOKS, M., COMITA, L., CRONE, E., CUBAYNES, S., DAVIES, T., VALPINE, P. et al. (2013). Strategies for fitting nonlinear ecological models in R, AD model builder, and BUGS. *Methods Ecol. Evol.* **4** 501–512.
- BREED, G. A., JONSEN, I. D., MYERS, R. A., BOWEN, W. D. and LEONARD, M. L. (2009). Sex-specific, seasonal foraging tactics of adult grey seals (*Halichoerus grypus*) revealed by state–space analysis. *Ecology* **90** 3209–3221.
- BREED, G. A., COSTA, D. P., JONSEN, I. D., ROBINSON, P. W. and MILLS-FLEMMING, J. (2012). State-space methods for more completely capturing behavioral dynamics from animal tracks. *Ecol. Model.* **235–236** 49–58.
- BRILLINGER, D. R., STEWART, B. S. and LITTNAN, C. L. (2008). Three months journeying of a Hawaiian monk seal. In *Probability and Statistics: Essays in Honor of David A. Freedman. Inst. Math. Stat. (IMS) Collect.* **2** 246–264. IMS, Beachwood, OH. [MR2459954](#)
- BRILLINGER, D. R., PREISLER, H. K., AGER, A. A., KIE, J. G. and STEWART, B. S. (2002). Employing stochastic differential equations to model wildlife motion. *Bull. Braz. Math. Soc. (N.S.)* **33** 385–408. [MR1978835](#)
- BUDERMAN, F. E., HOOTEN, M. B., IVAN, J. S. and SHENK, T. M. (2016). A functional model for characterizing long-distance movement behaviour. *Methods Ecol. Evol.* **7** 264–273.
- CARTON, J. A. and GIESE, B. S. (2008). A reanalysis of ocean climate using simple ocean data assimilation (SODA). *Mon. Weather Rev.* **136** 2999–3017.
- CARTON, J. A., CHEPURIN, G. A., CHEN, L. and GRODSKY, S. A. (2018). Improved global net surface heat flux. *J. Geophys. Res.* **123** 3144–3163.
- CONN, P. B., THORSON, J. T. and JOHNSON, D. S. (2017). Confronting preferential sampling when analysing population distributions: Diagnosis and model-based triage. *Methods Ecol. Evol.* **8** 1535–1546.
- DA SILVA FERREIRA, G. and GAMERMAN, D. (2015). Optimal design in geostatistics under preferential sampling. *Bayesian Anal.* **10** 711–735. [MR3420820](#)
- DIGGLE, P. J., MENEZES, R. and SU, T. (2010). Geostatistical inference under preferential sampling. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59** 191–232. [MR2744471](#)
- DIGGLE, P. J. and RIBEIRO, P. J. JR. (2007). *Model-Based Geostatistics. Springer Series in Statistics*. Springer, New York. [MR2293378](#)
- DINSDALE, D. and SALIBIAN-BARRERA, M. (2019a). Methods for preferential sampling in geostatistics. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 181–198. [MR3902988](#)
- DINSDALE, D. and SALIBIAN-BARRERA, M. (2019b). Supplement to “Modelling ocean temperatures from bio-probes under preferential sampling.” DOI:[10.1214/18-AOAS1217SUPP](#).
- DUJON, A. M., LINDSTROM, R. T. and HAYS, G. C. (2014). The accuracy of fastloc-GPS locations and implications for animal tracking. *Methods Ecol. Evol.* **5** 1162–1169.
- EVANS, K., LEA, M.-A. and PATTERSON, T. (2013). Recent advances in bio-logging science: Technologies and methods for understanding animal behaviour and physiology and their environments. *Deep-Sea Res., Part 2, Top. Stud. Oceanogr.* **88** 1–6.
- FEDAK, M. (2004). Marine animals as platforms for oceanographic sampling: A situation for biology and operational oceanography. *Mem. Natl. Inst. Polar Res., Spec. Issue* **58** 133–147.
- FEDAK, M. A. (2013). The impact of animal platforms on polar ocean observation. *Deep-Sea Res., Part 2, Top. Stud. Oceanogr.* **88–89** 7–13.
- GELFAND, A. E., SAHU, S. K. and HOLLAND, D. M. (2012). On the effect of preferential sampling in spatial prediction. *Environmetrics* **23** 565–578. [MR3020075](#)
- GLOAGUEN, P., ETIENNE, M.-P. and LE CORFF, S. (2018). Stochastic differential equation based on a multimodal potential to model movement data in ecology. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 599–619. [MR3787968](#)

- GNEITING, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19** 1327–1349. [MR3102554](#)
- GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. [MR2345548](#)
- GOULD, J., ROEMMICH, D., WIJFFELS, S., FREELAND, H., IGNASZEWSKY, M., JIANPING, X., POULIQUEN, S., DESAUBIES, Y., SEND, U., RADHAKRISHNAN, K. et al. (2004). Argo profiling floats bring new era of in situ ocean observations. *Eos Trans. AGU* **85** 185–191.
- GRIEWANK, A. and WALTHER, A. (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, 2nd ed. SIAM, Philadelphia, PA. [MR2454953](#)
- GUINET, C., VACQUIÉ-GARCIA, J., PICARD, B., BESSIGNEUL, G., LEBRAS, Y., DRAGON, A. C., VIVIAN, M., ARNOULD, J. P. and BAILLEUL, F. (2014). Southern elephant seal foraging success in relation to temperature and light conditions: Insight into prey distribution. *Mar. Ecol. Prog. Ser.* **499** 285–301.
- GURARIE, E. and OVASKAINEN, O. (2011). Characteristic spatial and temporal scales unify models of animal movement. *Amer. Nat.* **178** 113–123.
- GURARIE, E., FLEMING, C. H., FAGAN, W. F., LAIDRE, K. L., HERNÁNDEZ-PLIEGO, J. and OVASKAINEN, O. (2017). Correlated velocity models as a fundamental unit of animal movement: Synthesis and applications. *Mov. Eco.* **5** 13.
- HOOTEN, M. B. and JOHNSON, D. S. (2017a). Basis function models for animal movement. *J. Amer. Statist. Assoc.* **112** 578–589. [MR3671753](#)
- HOOTEN, M. B., JOHNSON, D. S., MCCLINTOCK, B. T. and MORALES, J. M. (2017b). *Animal Movement: Statistical Models for Telemetry Data*. CRC Press, Boca Raton, FL.
- JACOBS, S. (2006). Observations of change in the Southern Ocean. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **364** 1657–1681.
- JEONG, J. and JUN, M. (2015). A class of Matérn-like covariance functions for smooth processes on a sphere. *Spat. Stat.* **11** 1–18. [MR3311853](#)
- JOHNSON, D. S., LONDON, J. M., LEA, M.-A. and DURBAN, J. W. (2008). Continuous-time correlated random walk model for animal telemetry data. *Ecology* **89** 1208–1215.
- JONSEN, I. D., FLEMMING, J. M. and MYERS, R. A. (2005). Robust state-space modeling of animal movement data. *Ecology* **86** 2874–2880.
- KRISTENSEN, K., NIELSEN, A., BERG, C. W., SKAUG, H. and BELL, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *J. Stat. Softw.* **70** 1–21. DOI:[10.18637/jss.v070.i05](#).
- LINDGREN, F. and RUE, H. (2015). Bayesian spatial modelling with R-INLA. *J. Stat. Softw.* **63** 1–25.
- LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. [MR2853727](#)
- LIU, Y., BATTAILLE, B. C., TRITES, A. W. and ZIDEK, J. V. (2015). Bias correction and uncertainty characterization of dead-reckoned paths of marine mammals. *Animal Biotelemetry* **3** 51–61.
- MCCLINTOCK, B. T., KING, R., THOMAS, L., MATTHIOPOULOS, J., MCCONNELL, B. J. and MORALES, J. M. (2012). A general discrete-time modeling framework for animal movement using multistate random walks. *Ecol. Monogr.* **82** 335–349.
- MCINTYRE, T., ANSORGE, I. J., BORNEMANN, H., PLÖTZ, J., TOSH, C. A. and BESTER, M. N. (2011). Elephant seal dive behaviour is influenced by ocean temperature: Implications for climate change impacts on an ocean predator. *Mar. Ecol. Prog. Ser.* **441** 257–272.
- MORALES, J. M., HAYDON, D. T., FRAIR, J., HOLSINGER, K. E. and FRYXELL, J. M. (2004). Extracting more out of relocation data: Building movement models as mixtures of random walks. *Ecology* **85** 2436–2445.
- PATI, D., REICH, B. J. and DUNSON, D. B. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika* **98** 35–48. [MR2804208](#)

- PREISLER, H. K., AGER, A. A. and WISDOM, M. J. (2013). Analyzing animal movement patterns using potential functions. *Ecosphere* **4** 1–13.
- ROBUSTO, C. C. (1957). Classroom Notes: The Cosine-Haversine Formula. *Amer. Math. Monthly* **64** 38–40. [MR1529506](#)
- ROQUET, F., WUNSCH, C., FORGET, G., HEIMBACH, P., GUINET, C., REVERDIN, G., CHARRASSIN, J.-B., BAILLEUL, F., COSTA, D. P., HUCKSTADT, L. A. et al. (2013). Estimates of the Southern Ocean general circulation improved by animal-borne instruments. *Geophys. Res. Lett.* **40** 6176–6180.
- ROQUET, F., WILLIAMS, G., HINDELL, M. A., HARCOURT, R., MCMAHON, C., GUINET, C., CHARRASSIN, J.-B., REVERDIN, G., BOEHME, L., LOVELL, P. et al. (2014). A southern Indian Ocean database of hydrographic profiles obtained with instrumented elephant seals. *Sci. Data* **1** 140028.
- ROULSTON, M. S. and SMITH, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.* **130** 1653–1660.
- RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. [MR2649602](#)
- RUSSELL, J. C., HANKS, E. M., HARAN, M. and HUGHES, D. (2018). A spatially varying stochastic differential equation model for animal movement. *Ann. Appl. Stat.* **12** 1312–1331. [MR3834305](#)
- SHADDICK, G. and ZIDEK, J. V. (2014). A case study in preferential sampling: Long term monitoring of air pollution in the UK. *Spat. Stat.* **9** 51–65. [MR3326831](#)
- SIEGERT, S., FERRO, C. A. and STEPHENSON, D. B. (2014). Evaluating ensemble forecasts by the ignorance score—correcting the finite-ensemble bias. Preprint. Available at [arXiv:1410.8249](#).
- SIMPSON, D., LINDGREN, F. and RUE, H. (2012). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics* **23** 65–74. [MR2873784](#)
- UNGAR, E. D., HENKIN, Z., GUTMAN, M., DOLEV, A., GENIZI, A. and GANSKOPP, D. (2005). Inference of animal activity from GPS collar data on free-ranging cattle. *Rangeland Ecology and Management* **58** 256–266.
- VOTIER, S. C., BEARHOP, S., WITT, M. J., INGER, R., THOMPSON, D. and NEWTON, J. (2010). Individual responses of seabirds to commercial fisheries revealed using GPS tracking, stable isotopes and vessel monitoring systems. *J. Appl. Ecol.* **47** 487–497.
- WEIMERSKIRCH, H., BONADONNA, F., BAILLEUL, F., MABILLE, G., DELL’OMO, G. and LIPP, H.-P. (2002). GPS tracking of foraging albatrosses. *Science* **295** 1259.
- WHORISKEY, K., AUGER-MÉTHÉ, M., ALBERTSEN, C. M., WHORISKEY, F. G., BINDER, T. R., KRUEGER, C. C. and FLEMMING, J. M. (2017). A hidden Markov movement model for rapidly identifying behavioral states from animal tracks. *Ecol. Evol.* **7** 2112–2121.
- WILSON, R. and WILSON, M.-P. (1988). Dead reckoning: A new technique for determining penguin movements at sea. *Meeresforschung* **32** 155–158.

COMPLETE SPATIAL MODEL CALIBRATION¹

BY YEN-NING HUANG*, BRIAN J. REICH*, MONTSERRAT FUENTES[†] AND
 A. SANKARASUBRAMANIAN*

North Carolina State University and Virginia Commonwealth University[†]*

Computer simulation models are central to environmental science. These mathematical models are used to understand complex weather and climate patterns and to predict the climate’s response to different forcings. Climate models are of course not perfect reflections of reality, and so comparison with observed data is needed to quantify and to correct for biases and other deficiencies. We propose a new method to calibrate model output using observed data. Our approach not only matches the marginal distributions of the model output and gridded observed data, but it simultaneously postprocesses the model output to have the same spatial correlation as the observed data. This comprehensive calibration method permits realistic spatial simulations for regional impact studies. We apply the proposed method to global climate model output in North America and show that it successfully calibrates the model output for temperature and precipitation.

REFERENCES

- AZZALINI, A. and CAPITANIO, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t -distribution. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 367–389. [MR1983753](#)
- BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL.
- BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *J. Agric. Biol. Environ. Stat.* **15** 176–197. [MR2787270](#)
- BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2012). Space-time data fusion under error in computer model output: An application to modeling air quality. *Biometrics* **68** 837–848. [MR3055188](#)
- BERROCAL, V. J., RAFTERY, A. E. and GNEITING, T. (2007). Combining spatial statistical and ensemble information for probabilistic weather forecasting. *Mon. Weather Rev.* **135** 1386–1402.
- BERROCAL, V. J., RAFTERY, A. E. and GNEITING, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Ann. Appl. Stat.* **2** 1170–1193. [MR2655654](#)
- BHOWMIK, R. D., SANKARASUBRAMANIAN, A., SINHA, T., PATSKOSKI, J., MAHINTHAKUMAR, G. and KUNKEL, K. (2017). Multivariate downscaling approach preserving cross-correlations across climate variables for projecting hydrologic fluxes. *J. Hydrometeorol.* **18** 2187–2205.
- BRADLEY, J. R., WIKLE, C. K. and HOLAN, S. H. (2015). Spatio-temporal change of support with application to American Community Survey multi-year period estimates. *Stat* **4** 255–270. [MR3414659](#)

Key words and phrases. Bayesian methods, calibration, spatial statistics.

- BROOKS, S., GELMAN, A., JONES, G. L. and MENG, X.-L., eds. (2011). Handbook of Markov Chain Monte Carlo. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. [MR2742422](#)
- CANNON, A. J. (2016). Multivariate bias correction of climate model output: Matching marginal distributions and intervariable dependence structure. *J. Climate* **29** 7045–7064.
- CLARK, M., GANGOPADHYAY, S., HAY, L., RAJAGOPALAN, B. and WILBY, R. (2004). The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* **5** 243–262.
- CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. [MR1239641](#)
- CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatio-Temporal Data*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. [MR2848400](#)
- DETTINGER, M. D., CAYAN, D. R., MEYER, M. K. and JETON, A. E. (2004). Simulated hydrologic responses to climate variations and change in the Merced, Carson, and American river basins, Sierra Nevada, California, 1900–2099. *Clim. Change* **62** 283–317.
- DEVINENI, N. and SANKARASUBRAMANIAN, A. (2010). Improved categorical winter precipitation forecasts through multimodel combinations of coupled GCMs. *Geophys. Res. Lett.* **37** L24704.
- FELDMANN, K., SCHEUERER, M. and THORARINSDOTTIR, T. L. (2015). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Mon. Weather Rev.* **143** 955–971.
- FUENTES, M. and RAFTERY, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61** 36–45. [MR2129199](#)
- GEL, Y., RAFTERY, A. E. and GNEITING, T. (2004). Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation method. *J. Amer. Statist. Assoc.* **99** 575–583. [MR2086380](#)
- GOTWAY, C. A. and YOUNG, L. J. (2002). Combining incompatible spatial data. *J. Amer. Statist. Assoc.* **97** 632–648. [MR1951636](#)
- GUINNESS, J. and FUENTES, M. (2017). Circulant embedding of approximate covariances for inference from Gaussian data on large lattices. *J. Comput. Graph. Statist.* **26** 88–97. [MR3610410](#)
- HAY, L. E., WILBY, R. L. and LEAVESLEY, G. H. (2000). A comparison of delta change and downscaled GCM scenarios for three mountainous basins in the United States. *J. Am. Water Resour. Assoc.* **36** 387–397.
- JONES, M. C. and FADDY, M. J. (2003). A skew extension of the t -distribution, with applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 159–174. [MR1959820](#)
- LUERS, A. L., CAYAN, D. R., FRANCO, G., HANEMANN, M. and CROES, B. (2006). Our changing climate, assessing the risks to California. Report No. CEC-500-2006-077, 1–16. California Energy Commission Sacramento, CA.
- MAURER, E. P., WOOD, A. W., ADAM, J. C., LETTENMAIER, D. P. and NIJSSEN, B. (2002). A long-term hydrologically-based data set of land surface fluxes and states for the conterminous United States. *J. Climate* **15** 3237–3251.
- MCMILLAN, N. J., HOLLAND, D. M., MORARA, M. and FENG, J. (2010). Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics* **21** 48–65. [MR2842223](#)
- MÖLLER, A., LENKOSKI, A. and THORARINSDOTTIR, T. L. (2012). Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Q. J. R. Meteorol. Soc.* **139** 982–991.
- MORRIS, S. A., RIECH, B. J., THIBAUD, E. and COOLEY, D. (2017). A space-time skew- t model for threshold exceedances. *Biometrics* **73** 749–758. [MR3713109](#)

- O'BRIEN, T. P., SORNETTE, D. and MCPHERRO, R. L. (2001). Statistical asynchronous regression determining: The relationship between two quantities that are not measured simultaneously. *J. Geophys. Res.* **106** 13247–13259.
- REICH, B. J., CHANG, H. H. and FOLEY, K. M. (2014). A spectral method for spatial downscaling. *Biometrics* **70** 932–942. [MR3295754](#)
- SCHEFZIK, R. (2017). Ensemble calibration with preserved correlations: Unifying and comparing ensemble copula coupling and member-by-member postprocessing. *Q. J. R. Meteorol. Soc.* **143** 999–1008.
- SCHEFZIK, R., THORARINSDOTTIR, T. L. and GNEITING, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statist. Sci.* **28** 616–640. [MR3161590](#)
- SEO, S. B., SINHA, T., MAHINTHAKUMAR, G., SANKARASUBRAMANIAN, A. and KUMAR, M. (2016). Identification of dominant source of errors in developing streamflow and groundwater projection under near-term climate change. *J. Geophys. Res., Atmos.* **121** 7652–7672.
- STONER, A. M. K., HAYHOE, K., YANG, X. and WUEBBLES, D. J. (2012). An asynchronous regional regression model for statistical downscaling of daily climate variables. *Int. J. Climatol.* **33** 2473–2494.
- TAYLOR, K. E., STOUFFER, R. J. and MEEHL, G. A. (2012). An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93** 485–498.
- THORARINSDOTTIR, T. L., GNEITING, T. and GISSIBL, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA J. Uncertain. Quantificat.* **1** 522–534. [MR3283896](#)
- VAN VUUREN, D. P., EDMONDS, J., KAINUMA, M., RIAHI, K., THOMSON, A., HIBBARD, K., HURTT, G. C., KRAM, T., KREY, V., LAMARQUE, J.-F., MASUI, T., MEINSHAUSEN, M., NAKICENOVIC, N., SMITH, S. J. and ROSE, S. K. (2011). The representative concentration pathways: An overview. *Clim. Change* **109** 5–31.
- VRAC, M. and FRIEDERICH, P. (2015). Multivariate—intervariable, spatial, and temporal—bias correction. *J. Climate* **28** 218–237.
- WIKLE, C. K. and BERLINER, L. M. (2005). Combining information across spatial scales. *Technometrics* **47** 80–91. [MR2099410](#)
- WOOD, A. W., LEUNG, L. R., SRIDHAR, V. and LETTENMAIER, D. P. (2004). Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Clim. Change* **62** 189–216.
- YOUNG, L. J. and GOTWAY, C. A. (2007). Linking spatial data from different sources: The effects of change of support. *Stoch. Environ. Res. Risk Assess.* **21** 589–600. [MR2380677](#)
- ZIDEK, J. V., LE, N. D. and LIU, Z. (2012). Combining data and simulated data for space-time fields: Application to ozone. *Environ. Ecol. Stat.* **19** 37–56. [MR2909084](#)

EXTENDED SENSITIVITY ANALYSIS FOR HETEROGENEOUS UNMEASURED CONFOUNDING WITH AN APPLICATION TO SIBLING STUDIES OF RETURNS TO EDUCATION

BY COLIN B. FOGARTY¹ AND RAIDEN B. HASEGAWA¹

Massachusetts Institute of Technology and University of Pennsylvania

The conventional model for assessing insensitivity to hidden bias in paired observational studies constructs a worst-case distribution for treatment assignments subject to bounds on the maximal bias to which any given pair is subjected. In studies where rare cases of extreme hidden bias are suspected, the maximal bias may be substantially larger than the typical bias across pairs, such that a correctly specified bound on the maximal bias would yield an unduly pessimistic perception of the study's robustness to hidden bias. We present an extended sensitivity analysis which allows researchers to simultaneously bound the maximal and typical bias perturbing the pairs under investigation while maintaining the desired Type I error rate. We motivate and illustrate our method with two sibling studies on the impact of schooling on earnings, one containing information of cognitive ability of siblings and the other not. Cognitive ability, clearly influential of both earnings and degree of schooling, is likely similar between members of most sibling pairs yet could, conceivably, vary drastically for some siblings. The method is straightforward to implement, simply requiring the solution to a quadratic program. R code is provided in the Supplementary Material.

REFERENCES

- ASHENFELTER, O. and ROUSE, C. (1998). Income, schooling, and ability: Evidence from a new sample of identical twins. *Q. J. Econ.* **113** 253–284.
- BECKER, G. S. (2009). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Univ. Chicago Press, Chicago, IL.
- BERGER, R. L. and BOOS, D. D. (1994). *P* values maximized over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* **89** 1012–1016. [MR1294746](#)
- BHATIA, R. and DAVIS, C. (2000). A better bound on the variance. *Amer. Math. Monthly* **107** 353–357. [MR1763060](#)
- CARD, D. (1999). The causal effect of education on earnings. In *Handbook of Labor Economics* (O. C. Ashenfelter and D. Card, eds.) **3** 1801–1863. Elsevier, Amsterdam.
- CARD, D. and KRUEGER, A. B. (1992). Does school quality matter? Returns to education and the characteristics of public schools in the United States. *J. Polit. Econ.* **100** 1–40.
- CORNFIELD, J., HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A. M., SHIMKIN, M. B. and WYNDER, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* **22** 173–203.
- DONOVAN, S. J. and SUSSER, E. (2011). Commentary: Advent of sibling designs. *Int. J. Epidemiol.* **40** 345–349.

Key words and phrases. Observational studies, causal inference, nuisance parameters, quadratic programming, hidden bias, superpopulation inference.

- EGLESTON, B. L., SCHARFSTEIN, D. O. and MACKENZIE, E. (2009). On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics* **65** 497–504. [MR2751473](#)
- FISHER, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.
- FOGARTY, C. B. and HASEGAWA, R. B. (2019). Supplement to “Extended sensitivity analysis for heterogeneous unmeasured confounding with an application to sibling studies of returns to education.” DOI:[10.1214/18-AOAS1215SUPPA](#), DOI:[10.1214/18-AOAS1215SUPPB](#).
- FOGARTY, C. B. and SMALL, D. S. (2016). Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *J. Amer. Statist. Assoc.* **111** 1820–1830. [MR3601738](#)
- FRISELL, T. (2012). Sibling comparison designs: Bias from non-shared confounders and measurement error. *Epidemiology* **23** 713–720.
- GRILICHES, Z. (1970). Notes on the role of education in production functions and growth accounting. In *Education, Income, and Human Capital* NBER Chapters 71–127. National Bureau of Economic Research, Cambridge, MA.
- GRILICHES, Z. (1979). Sibling models and data in economics: Beginnings of a survey. *J. Polit. Econ.* **87** S37–S64.
- HÁJEK, J., ŠIDÁK, Z. and SEN, P. K. (1999). *Theory of Rank Tests*, 2nd ed. *Probability and Mathematical Statistics*. Academic Press, San Diego, CA. [MR1680991](#)
- HANSEN, B. B. and KLOPPER, S. O. (2006). Optimal full matching and related designs via network flows. *J. Comput. Graph. Statist.* **15** 609–627. [MR2280151](#)
- HASEGAWA, R. and SMALL, D. (2017). Sensitivity analysis for matched pair analysis of binary data: From worst case to average case analysis. *Biometrics* **73** 1424–1432. [MR3744554](#)
- HAUSER, R. M., SHERIDAN, J. T., WARREN and ROBERT, J. (1999). Socioeconomic achievements of siblings in the life course. *Res. Aging* **21** 338–378. DOI:[10.1177/0164027599212008](#).
- HERD, P., CARR, D. and ROAN, C. (2014). Cohort profile: Wisconsin longitudinal study (WLS). *Int. J. Epidemiol.* **43** 34–41.
- HOSMAN, C. A., HANSEN, B. B. and HOLLAND, P. W. (2010). The sensitivity of linear regression coefficients’ confidence limits to the omission of a confounder. *Ann. Appl. Stat.* **4** 849–870. [MR2758424](#)
- HSU, J. Y. and SMALL, D. S. (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics* **69** 803–811. [MR3146776](#)
- IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* **93** 126–132.
- LIU, W., KURAMOTO, JANET, S. and STUART, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prev. Sci.* **14** 570–580.
- MARCUS, S. M. (1997). Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *J. Educ. Behav. Stat.* **22** 193–201.
- ROSENBAUM, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *J. Amer. Statist. Assoc.* **79** 565–574. [MR0763575](#)
- ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. [MR0885915](#)
- ROSENBAUM, P. R. (1992). Detecting bias with confidence in observational studies. *Biometrika* **79** 367–374.
- ROSENBAUM, P. R. (1999). Choice as an alternative to control in observational studies. *Statist. Sci.* **14** 259–278.
- ROSENBAUM, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. [MR1962487](#)
- ROSENBAUM, P. R. (2002b). *Observational Studies*, 2nd ed. Springer, New York.
- ROSENBAUM, P. R. (2010). *Design of Observational Studies*. *Springer Series in Statistics*. Springer, New York. [MR2561612](#)

- ROSENBAUM, P. R. and SILBER, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *J. Amer. Statist. Assoc.* **104** 1398–1405. [MR2750570](#)
- RUBIN, D. B. (1980). Comment on D. Basu, “Randomization analysis of experimental data: The Fisher randomization test.” *J. Amer. Statist. Assoc.* **75** 591–593.
- STANEK, K. C., IACONO, WILLIAM, G. and MCGUE, M. (2011). Returns to education: What do twin studies control? *Twin Res. Hum. Genet.* **14** 509–515.
- STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. [MR2741812](#)
- VANDERWEELE, T. J. and DING, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Ann. Intern. Med.* **167** 268–274.
- WANG, L. and KRIEGER, A. M. (2006). Causal conclusions are most sensitive to unobserved binary covariates. *Stat. Med.* **25** 2257–2271. [MR2240099](#)
- YU, B. and GASTWIRTH, J. L. (2005). Sensitivity analysis for trend tests: Application to the risk of radiation exposure. *Biostatistics* **6** 201–209.
- ZUBIZARRETA, J. R., CERDÁ, M. and ROSENBAUM, P. R. (2013). Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design. *Epidemiology* **24** 79–87.

A HIERARCHICAL MULTIVARIATE SPATIO-TEMPORAL MODEL FOR CLUSTERED CLIMATE DATA WITH ANNUAL CYCLES

BY GIANLUCA MASTRANTONIO^{*,1}, GIOVANNA JONA LASINIO^{†,1},
 ALESSIO POLLICE^{‡,1}, GIULIA CAPOTORTI[†], LORENZO TEODONIO[§],
 GIULIO GENOVA[†] AND CARLO BLASI[†]

Politecnico di Torino^{}, Sapienza Università di Roma[†], Università degli
 Studi di Bari “Aldo Moro”[‡] and Ministry of Cultural Heritage and
 Activities and Tourism[§]*

We introduce a Bayesian multivariate hierarchical framework to estimate a space-time model for a joint series of monthly extreme temperatures and amounts of precipitation. Data are available for 360 monitoring stations over 60 years, with missing data affecting almost all series. Model components account for spatio-temporal correlation and annual cycles, dependence on covariates and between responses. Spatio-temporal dependence is modeled by the nearest neighbor Gaussian process (GP), response multivariate dependencies are represented by the linear model of coregionalization and effects of annual cycles are included by a circular representation of time. The proposed approach allows imputation of missing values and interpolation of climate surfaces at the national level. It also provides a characterization of the so called Italian ecoregions, namely broad and discrete ecologically homogeneous areas of similar potential as regards the climate, physiography, hydrography, vegetation and wildlife. To now, Italian ecoregions are hierarchically classified into 4 tiers that go from 2 Divisions to 35 Subsections and are defined by informed expert judgments. The current climatic characterization of Italian ecoregions is based on bioclimatic indices for the period 1955–2000.

REFERENCES

- ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* **18** 343–373. [MR2461882](#)
- BAI, Y., SONG, P. X.-K. and RAGHUNATHAN, T. E. (2012). Joint composite estimating functions in spatiotemporal models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 799–824. [MR2988907](#)
- BAILEY, R. G. (1983). Delineation of ecosystem regions. *Environ. Manag.* **7** 365–373.
- BAILEY, R. G. (2004). Identifying ecoregion boundaries. *Environ. Manag.* **34** S14–S26.
- BALINT, P. J., STEWART, R. E., DESAI, A. and WALTERS, L. C. (2011). *Wicked Environmental Problems*. Island Press, Washington, DC.
- BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. *Monographs on Statistics and Applied Probability* **135**. CRC Press, Boca Raton, FL. [MR3362184](#)
- BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 825–848. [MR2523906](#)

Key words and phrases. Multivariate process, cyclic effect, coregionalization, NNGP.

- BEVILACQUA, M., GAETAN, C., MATEU, J. and PORCU, E. (2012). Estimating space and space-time covariance functions for large data sets: A weighted composite likelihood approach. *J. Amer. Statist. Assoc.* **107** 268–280. [MR2949358](#)
- BEVILACQUA, M., FASSÒ, A., GAETAN, C., PORCU, E. and VELANDIA, D. (2016a). Covariance tapering for multivariate Gaussian random fields estimation. *Stat. Methods Appl.* **25** 21–37. [MR3460484](#)
- BEVILACQUA, M., ALEGRIA, A., VELANDIA, D. and PORCU, E. (2016b). Composite likelihood inference for multivariate Gaussian random fields. *J. Agric. Biol. Environ. Stat.* **21** 448–469. [MR3542081](#)
- BLANGIARDO, M. and CAMELETTI, M. (2015). *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. Wiley, Chichester. [MR3364017](#)
- BLASI, C., CAPOTORTI, G., COPIZ, R., GUIDA, D., MOLLO, B., SMIRAGLIA, D. and ZAVAT-TERO, L. (2014). Classification and mapping of the ecoregions of Italy. *Plant Biosyst.* **148** 1255–1345.
- CHUNG, M., BINOIS, M., GRAMACY, R. B., MOQUIN, D. J., SMITH, A. P. and SMITH, A. M. (2018). Parameter and uncertainty estimation for dynamical systems using surrogate stochastic processes. Preprint. Available at [arXiv:1802.00852](#).
- DANIELS, M. J. and KASS, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *J. Amer. Statist. Assoc.* **94** 1254–1263. [MR1731487](#)
- DANIELS, M. J. and POURAHMADI, M. (2009). Modeling covariance matrices via partial autocorrelations. *J. Multivariate Anal.* **100** 2352–2363. [MR2560376](#)
- DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016a). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. [MR3538706](#)
- DATTA, A., BANERJEE, S., FINLEY, A. O., HAMM, N. A. S. and SCHAAP, M. (2016b). Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *Ann. Appl. Stat.* **10** 1286–1316. [MR3553225](#)
- DIGGLE, P. J. and RIBEIRO, P. J. JR. (2002). Bayesian inference in Gaussian model-based geostatistics. *Geogr. Environ. Model.* **6** 129–146.
- EIDSVIK, J., SHABY, B. A., REICH, B. J., WHEELER, M. and NIEMI, J. (2014). Estimation and prediction in spatial models with block composite likelihoods. *J. Comput. Graph. Statist.* **23** 295–315. [MR3215812](#)
- FAYE, E., HERRERA, M., BELLOMO, L., SILVAIN, J.-F. and DANGLES, O. (2014). Strong discrepancies between local temperature mapping and interpolated climatic grids in tropical mountainous agricultural landscapes. *PLoS ONE* **9** e105541.
- FERRARO PETRILLO, U. and RAIMATO, G. (2014). TeraStat Computer cluster for high performance computing, Dept. Statistical Science, Sapienza Univ. Rome. Available at <http://www.dss.uniroma1.it/en/node/6554>.
- FICK, S. E. and HIJMANS, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37** 4302–4315.
- FINLEY, A. O., BANERJEE, S. and GELFAND, A. E. (2012). Bayesian dynamic modeling for large space-time datasets using Gaussian predictive processes. *J. Geogr. Syst.* **14** 29–47.
- FINLEY, A. O., DATTA, A., COOK, B. C., MORTON, D. C., ANDERSEN, H. E. and BANERJEE, S. (2017). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. Preprint. Available at [arXiv:1702.00434](#).
- GELFAND, A. E., SCHMIDT, A. M., BANERJEE, S. and SIRMANS, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST* **13** 263–312. [MR2154003](#)
- GELFAND, A. E., DIGGLE, P. J., FUENTES, M. and GUTTORP, P., eds. (2010). *Handbook of Spatial Statistics. Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. [MR2761512](#)

- GNEITING, T. (2002). Nonseparable, stationary covariance functions for space-time data. *J. Amer. Statist. Assoc.* **97** 590–600. [MR1941475](#)
- GNEITING, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19** 1327–1349. [MR3102554](#)
- GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. [MR2345548](#)
- GRÄLER, B., PEBESMA, E. and HEUVELINK, G. (2016). Spatio-temporal interpolation using gstat. *RFID J.* **8** 204–218.
- GRAMACY, R. B. and APLEY, D. W. (2015). Local Gaussian process approximation for large computer experiments. *J. Comput. Graph. Statist.* **24** 561–578. [MR3357395](#)
- GRIMIT, E. P., GNEITING, T., BERROCAL, V. J. and JOHNSON, N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Q. J. R. Meteorol. Soc.* **132** 2925–2942.
- GUINNESS, J. (2018). Permutation and grouping methods for sharpening Gaussian process approximations. *Technometrics*. To appear.
- HANNAH, L., ROEHRDANZ, P. R., IKEGAMI, M., SHEPARD, A. V., SHAW, M. R., TABOR, G., ZHI, L., MARQUET, P. A. and HIJMANS, R. J. (2013). Climate change, wine, and conservation. *Proc. Natl. Acad. Sci. USA* **110** 6907–6912.
- HEATON, M. J., DATTA, A., FINLEY, A., FURRER, R., GUHANIYOGI, R., GERBER, F., GRAMACY, R. B., HAMMERLING, D., KATZFUSS, M., LINDGREN, F., NYCHKA, D. W., SUN, F. and ZAMMIT-MANGION, A. (2017). A case study competition among methods for analyzing large spatial data. Preprint. Available at [arXiv:1710.05013](#).
- HIJMANS, R. J., CAMERON, S. E., PARRA, J. L., JONES, P. G. and JARVIS, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25** 1965–1978.
- JONA LASINIO, G., MASTRANTONIO, G. and POLLICE, A. (2013). Discussing the “big n problem.” *Stat. Methods Appl.* **22** 97–112. [MR3038184](#)
- KATZFUSS, M. and GUINNESS, J. (2017). A general framework for Vecchia approximations of Gaussian processes. Preprint. Available at [arXiv:1708.06302](#).
- LI, S., DRAGICEVIC, S., CASTRO, F. A., SESTER, M., WINTER, S., COLTEKIN, A., PETTIT, C., JIANG, B., HAWORTH, J., STEIN, A. and CHENG, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS J. Photogramm. Remote Sens.* **115** 119–133.
- LOVELAND, T. R. and MERCHANT, J. M. (2004). Ecoregions and ecoregionalization: Geographical and ecological perspectives. *Environ. Manag.* **34** S1.
- METZGER, M. J., BUNCE, R. G. H., JONGMAN, R. H. G., SAYRE, R., TRABUCCO, A. and ZOMER, R. (2013). A high-resolution bioclimate map of the world: A unifying framework for global biodiversity research and monitoring. *Glob. Ecol. Biogeogr.* **22** 630–638.
- MÜCHER, C. A., KLIJN, J. A., WASCHER, D. M. and SCHAMINÉE, J. H. J. (2010). A new European landscape classification (LANMAP): A transparent, flexible and user-oriented methodology to distinguish landscapes. *Ecol. Indicators* **10** 87–103.
- NYCHKA, D., FURRER, R., PAIGE, J. and SAIN, S. (2017). fields: Tools for spatial data. R package version 9.6, Univ. Corporation for Atmospheric Research, Boulder, CO.
- PEBESMA, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Comput. Geosci.* **30** 683–691.
- PESARESI, S., GALDENZI, D., BIONDI, E. and CASAVECCHIA, S. (2014). Bioclimate of Italy: Application of the worldwide bioclimatic classification system. *J. Maps* **10** 538–553.
- R CORE TEAM (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- SHIROTA, S. and GELFAND, A. E. (2017). Space and circular time log Gaussian Cox processes with application to crime event data. *Ann. Appl. Stat.* **11** 481–503. [MR3693535](#)

- SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. [MR1979380](#)
- WMO (2016). *Guidelines on Best Practices for Climate Data Rescue*. WMO—Guidelines **1182**. World Meteorological Organization, Geneva.
- WOOD, S. N. (2017). *Generalized Additive Models: An Introduction with R*, 2nd ed. *Texts in Statistical Science Series*. CRC Press/CRC, Boca Raton, FL. [MR2206355](#)
- XU, G., LIANG, F. and GENTON, M. G. (2015). A Bayesian spatio-temporal geostatistical model with an auxiliary lattice for large datasets. *Statist. Sinica* **25** 61–79. [MR3328803](#)

FUSED COMPARATIVE INTERVENTION SCORING FOR HETEROGENEITY OF LONGITUDINAL INTERVENTION EFFECTS¹

BY JARED D. HULING^{*}, MENGANG YU[†] AND MAUREEN SMITH[†]

Ohio State University^{} and University of Wisconsin-Madison[†]*

With the growing cost of health care in the United States, the need to improve efficiency and efficacy has become increasingly urgent. There has been a keen interest in developing interventions to effectively coordinate the typically fragmented care of patients with many comorbidities. Evaluation of such interventions is often challenging given their long-term nature and their differential effectiveness among different patients. Furthermore, care coordination interventions are often highly resource-intensive. Hence there is pressing need to identify which patients would benefit the most from a care coordination program. In this work we introduce a subgroup identification procedure for long-term interventions whose effects are expected to change smoothly over time. We allow differential effects of an intervention to vary over time and encourage these effects to be more similar for closer time points by utilizing a fused lasso penalty. Our approach allows for flexible modeling of temporally changing intervention effects while also borrowing strength in estimation over time. We utilize our approach to construct a personalized enrollment decision rule for a complex case management intervention in a large health system and demonstrate that the enrollment decision rule results in improvement in health outcomes and care costs. The proposed methodology could have broad usage for the analysis of different types of long-term interventions or treatments including other interventions commonly implemented in health systems.

REFERENCES

- BLUMENTHAL, D., ANDERSON, G., BURKE, S., et al. (2016). Tailoring complex-care management, coordination, and integration for high need, high cost patients. Technical report, Discussion paper, National Academy of Medicine, Washington, DC.
- BODENHEIMER, T. and BERRY-MILLET, R. (2009). Care management of patients with complex health care needs. In *The Synthesis Project*. Research synthesis report 19. Robert Wood Johnson Foundation, Princeton, NJ.
- BOYD, S., PARIKH, N., CHU, E., PELEATO, B., ECKSTEIN, J.. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.
- CHEN, S., TIAN, L., CAI, T. and YU, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* **73** 1199–1209. [MR3744534](#)

Key words and phrases. Fused lasso, precision medicine, comparative effectiveness research, electronic health records, interaction modeling.

- CHEUNG, W. Y., NEVILLE, B. A., CAMERON, D. B., COOK, E. F. and EARLE, C. C. (2009). Comparisons of patient and physician expectations for cancer survivorship care. *J. Clin. Oncol.* **27** 2489–2495.
- COHEN, S. B. and YU, W. (2012). Statistical brief # 354 2008–2009. Agency for Healthcare Research and Quality, Washington, DC.
- EFRON, B. (2014). Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.* **109** 991–1007. [MR3265671](#)
- FAN, J. and ZHANG, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518. [MR1742497](#)
- FAN, J. and ZHANG, W. (2008). Statistical methods with varying coefficient models. *Stat. Interface* **1** 179–195. [MR2425354](#)
- FOSTER, J. C., TAYLOR, J. M. G. and RUBERG, S. J. (2011). Subgroup identification from randomized clinical trial data. *Stat. Med.* **30** 2867–2880. [MR2844689](#)
- HARRELL, F. E., LEE, K. L. and MARK, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15** 361–387.
- HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. [MR1229881](#)
- HICKHAM, D., WEISS, J. W., GUISE, J., BUCKLEY, D., MOTU'APUAKA, M., GRAHAM, E., WASSON, N. and SAHA, S. (2013). Case management for adults with medical illness and complex care needs. Comparative Effectiveness Reviews No. 99, AHRQ Publication No. 13-EHC031-EF. Accessed on November 13, 2015. Available at <http://www.ncbi.nlm.nih.gov/pubmed/23346604>.
- HULING, J. D., YU, M. and SMITH, M. (2019). Supplement to “Fused comparative intervention scoring for heterogeneity of longitudinal intervention effects.” DOI:[10.1214/18-AOAS1216SUPPA](https://doi.org/10.1214/18-AOAS1216SUPPA), DOI:[10.1214/18-AOAS1216SUPPB](https://doi.org/10.1214/18-AOAS1216SUPPB).
- IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—For Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. [MR3309951](#)
- POPE, G. C., ELLIS, R. P., ASH, A. S., AYANIAN, J. Z., BATES, D. W., BURSTIN, H., IEZZONI, L. I., MARCANTONIO, E. and WU, B. (2000). Diagnostic cost group hierarchical condition category models for Medicare risk adjustment. Health Economics Research, Inc., Waltham, MA.
- QIAN, M. and MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210. [MR2816351](#)
- ROBINS, J. M. and FINKELSTEIN, D. M. (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* **56** 779–788.
- ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. [MR0742974](#)
- RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. [MR2166071](#)
- SCHNEEWEISS, S., RASSEN, J. A., GLYNN, R. J., AVORN, J., MOGUN, H. and BROOKHART, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20** 512.
- STANGE, K. C. (2009). The problem of fragmentation and the need for integrative solutions. *Ann. Fam. Med.* **7** 100–103.
- STANGE, K. C. (2012). In this issue: Challenges of managing multimorbidity. *Ann. Fam. Med.* **10** 2–3.
- TIAN, L., ALIZADEH, A. A., GENTLES, A. J. and TIBSHIRANI, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *J. Amer. Statist. Assoc.* **109** 1517–1532. [MR3293607](#)

- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. [MR2850205](#)
- TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. [MR2136641](#)
- U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES (2012). *Multiple Chronic Conditions Initiative. Private Sector Activities Focused on Improving the Health of Individuals with Multiple Chronic Conditions: Innovative Profiles*. U.S. Dept. Health and Human Services, Washington, DC. Available at <http://www.hhs.gov/ash/initiatives/mcc/>.
- ZHOU, X., KOSOROK, M. R. (2017). Augmented outcome-weighted learning for optimal treatment regimes. Preprint. Available at [arXiv:1711.10654](#).
- ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** 1106–1118. [MR3010898](#)
- ZHOU, X., MAYER-HAMBLETT, N., KHAN, U. and KOSOROK, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *J. Amer. Statist. Assoc.* **112** 169–187. [MR3646564](#)

GRAPHICAL MODELS FOR ZERO-INFLATED SINGLE CELL GENE EXPRESSION

BY ANDREW MCDAVID^{*,1}, RAPHAEL GOTTARDO^{†,‡,1,2}, NOAH SIMON[‡] AND
MATHIAS DRTON^{‡,§,3}

University of Rochester Medical Center^{}, Fred Hutchinson Cancer Research Center[†], University of Washington[‡] and University of Copenhagen[§]*

Bulk gene expression experiments relied on aggregations of thousands of cells to measure the average expression in an organism. Advances in microfluidic and droplet sequencing now permit expression profiling in single cells. This study of cell-to-cell variation reveals that individual cells lack detectable expression of transcripts that appear abundant on a population level, giving rise to zero-inflated expression patterns. To infer gene coregulatory networks from such data, we propose a multivariate Hurdle model. It is comprised of a mixture of singular Gaussian distributions. We employ neighborhood selection with the pseudo-likelihood and a group lasso penalty to select and fit undirected graphical models that capture conditional independences between genes. The proposed method is more sensitive than existing approaches in simulations, even under departures from our Hurdle model. The method is applied to data for T follicular helper cells, and a high-dimensional profile of mouse dendritic cells. It infers network structure not revealed by other methods, or in bulk data sets. A R implementation is available at <https://github.com/amcdavid/HurdleNormal>.

REFERENCES

- ADACHI, Y., HIRAMATSU, S., TOKUDA, N., SHARIFI, K., EBRAHIMI, M., ISLAM, A., KAGAWA, Y., KOSHY VAIDYAN, L., SAWADA, T., HAMANO, K. and OWADA, Y. (2012). Fatty acid-binding protein 4 (FABP4) and FABP5 modulate cytokine production in the mouse thymic epithelial cells. *Histochem. Cell Biol.* **138** 397–406.
- CHEN, S., WITTEN, D. M. and SHOJAIE, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102** 47–64. [MR3335095](#)
- CHENG, J., LI, T., LEVINA, E. and ZHU, J. (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Statist.* **26** 367–378. [MR3640193](#)
- THE GENE ONTOLOGY CONSORTIUM Gene ontology consortium: Going forward. *Nucleic Acids Res.* **43**. (D1): D1049–D1056, 2015.
- DE JONG, E. C., VIEIRA, P. L., KALINSKI, P., SCHUITEMAKER, J. H. N., TANAKA, Y., WIERENGA, E. A., YAZDANBAKHSH, M. and KAPSENBERG, M. L. (2002). Microbial compounds selectively induce Th1 cell-promoting or Th2 cell-promoting dendritic cells in vitro with diverse th cell-polarizing signals. *J. Immunol.* **168** 1704–1709.
- DENDA-NAGAI, K., AIDA, S., SABA, K., SUZUKI, K., MORIYAMA, S., OO-PUTHINAN, S., TSUIJI, M., MORIKAWA, A., KUMAMOTO, Y., SUGIURA, D., KUDO, A., AKIMOTO, Y., KAWAKAMI, H., BOVIN, N. V. and IRIMURA, T. (2010). Distribution and function of

Key words and phrases. Gene network, single cell gene expression, graphical model, group lasso.

- macrophage galactose-type C-type lectin 2 (MGL2/CD301b): Efficient uptake and presentation of glycosylated antigens by dendritic cells. *J. Biol. Chem.* **285** 19193–19204.
- DOBRA, A., HANS, C., JONES, B., NEVINS, J. R., YAO, G. and WEST, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.* **90** 196–212. [MR2064941](#)
- DRTON, M. and MAATHUIS, M. (2017). Structure learning in graphical modeling. *Annu. Rev. Stat. Appl.* **4** 365–393.
- DRTON, M., STURMFELS, B. and SULLIVANT, S. (2009). *Lectures on Algebraic Statistics. Oberwolfach Seminars* **39**. Birkhäuser, Basel. [MR2723140](#)
- ELTOFT, T., KIM, T. and LEE, T. W. (2006). On the multivariate Laplace distribution. *IEEE Signal Process. Lett.* **13** 300–303.
- FINAK, G., MCDAVID, A., YAJIMA, M., DENG, J., GERSUK, V., SHALEK, A. K., SLICHTER, C. K., MILLER, H. W., JULIANA MCEL RATH, M., PR LIC, M., LINSLEY, P. S. and GOTTARDO, R. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16** 278.
- FOYGEL, R. and DRTON, M. (2010). Exact block-wise optimization in group lasso and sparse group lasso for linear regression. 1–19. Arxiv preprint. Available at [arXiv:1010.3320](#).
- MARINOV, G. K., WILLIAMS, B. A., MCCUE, K., SCHROTH, G. P., GERTZ, J., MYERS, R. M. and WOLD, B. J. (2014). From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.* **24** 496–510.
- HERMANN-KLEITER, N. and BAIER, G. (2010). NFAT pulls the strings during CD4+ T helper cell effector functions. Unpublished manuscript.
- JANES, K. A., WANG, C.-C., HOLMBERG, K. J., CABRAL, K. and BRUGGE, J. S. (2010). Identifying single-cell molecular programs by stochastic profiling. *Nat. Methods* **7** 311–317.
- JOHNSTON, R. J., POHOLEK, A. C., DITORO, D., YUSUF, I., ETO, D., BARNETT, B., DENT, A. L., CRAFT, J. and CROTTY, S. (2009). Bcl6 and Blimp-1 are reciprocal and antagonistic regulators of T follicular helper cell differentiation. *Science* **325**.
- KIM, J. K. and MARIONI, J. C. (2013). Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.* **14**.
- PHAM, L. V., TAMAYO, A. T., YOSHIMURA, L. C., LIN-LEE, Y. C. and FORD, R. J. (2005). Constitutive NF-kappaB and NFAT activation in aggressive B-cell lymphomas synergistically activates the CD154 gene and maintains lymphoma cell survival. *Blood* **106** 3940–3947.
- LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. Oxford University Press, New York. [MR1419991](#)
- LEE, J. D. and HASTIE, T. J. (2013). Structure learning of mixed graphical models. In *AISTATS 16* **31** 388–396, Scottsdale, AZ. Available at <http://jmlr.org/proceedings/papers/v31/lee13a.html>.
- LI, Y., PEARL, S. A. and JACKSON, S. A. (2015). Gene networks in plant biology: Approaches in reconstruction and analysis. *Trends Plant Sci.* **20** 664–675.
- LIN, L., FINAK, G., USHEY, K., SESHADRI, C., HAWN, T. R., FRAHM, N., SCRIBA, T. J., MAHOMED, H., HANEKOM, W. et al. (2015). COMPASS identifies T-cell subsets correlated with clinical outcomes. *Nat. Biotechnol.* **33** 610–616.
- MA, C. S., DEENICK, E. K., BATTEN, M. and TANGYE, S. G. (2012). The origins, function, and regulation of T follicular helper cells. *J. Exp. Med.* **209** 1241–1253.
- MARKOWETZ, F. and SPANG, R. (2007). Inferring cellular networks: A review. *BMC Bioinform.* **8**.
- MCDAVID, A., FINAK, G., CHATTOPADYAY, P. K., DOMINGUEZ, M., LAMOREAUX, L., MA, S. S., ROEDERER, M. and GOTTARDO, R. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29** 461–467.
- MCDAVID, A., GOTTARDO, R., SIMON, N. and DRTON, M. (2019). Supplement to “Graphical models for zero-inflated single cell gene expression.” DOI:10.1214/18-AOAS1213SUPP.
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. [MR2278363](#)

- PARIKH, N. and BOYD, S. (2014). Proximal algorithms. *Found. Trends Optim.* **1** 123–231.
- PRECOPIO, M. L., BETTS, M. R., PARRINO, J., PRICE, D. A., GOSTICK, E., AMBROZAK, D. R., ASHER, T. E., DOUEK, D. C., HARARI, A. et al. (2007). Immunization with vaccinia virus induces polyfunctional and phenotypically distinctive CD8(+) T cell responses. *J. Exp. Med.* **204** 1405–1416.
- RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1 -regularized logistic regression. *Ann. Statist.* **38** 1287–1319. [MR2662343](#)
- SHAH, R. D. and SAMWORTH, R. J. (2013). Variable selection with error control: Another look at stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 55–80. [MR3008271](#)
- SHALEK, A. K., SATIJA, R., SHUGA, J., TROMBETTA, J. J., GENNERT, D., LU, D., CHEN, P., GERTNER, R. S., GAUBLomme, J. T. et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510** 263–269.
- SIMON, N. and TIBSHIRANI, R. (2012). Standardization and the group Lasso penalty. *Statist. Sinica* **22** 983–1001.
- TANSEY, W., PADILLA, O. H. M., SUGGALA, A. S. and RAVIKUMAR, P. (2015). Vector-space Markov random fields via exponential families. In *Proceedings of the 32nd International Conference on Machine Learning* **37** 684–692. Available at <http://jmlr.org/proceedings/papers/v37/tansey15.html>.
- TIBSHIRANI, R., BIEN, J., FRIEDMAN, J., HASTIE, T., SIMON, N., TAYLOR, J. and TIBSHIRANI, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 245–266. [MR2899862](#)
- YANG, E., BAKER, Y., RAVIKUMAR, P., ALLEN, G. and LIU, Z. (2014). Mixed graphical models via exponential families. In *AISTATS 17* **33**. Reykjavik, Iceland.

TREECLONE: RECONSTRUCTION OF TUMOR SUBCLONE PHYLOGENY BASED ON MUTATION PAIRS USING NEXT GENERATION SEQUENCING DATA¹

BY TIANJIAN ZHOU^{*,†,‡,2}, SUBHAJIT SENGUPTA^{†,2}, PETER MÜLLER[‡] AND
YUAN JI^{*,†}

The University of Chicago^{}, NorthShore University HealthSystem[†] and
The University of Texas at Austin[‡]*

We present TreeClone, a latent feature allocation model to reconstruct tumor subclones subject to phylogenetic evolution that mimics tumor evolution. Similar to most current methods, we consider data from next-generation sequencing of tumor DNA. Unlike most methods that use information in short reads mapped to single nucleotide variants (SNVs), we consider subclone phylogeny reconstruction using pairs of two proximal SNVs that can be mapped by the same short reads. As part of the Bayesian inference model, we construct a phylogenetic tree prior. The use of the tree structure in the prior greatly strengthens inference. Only subclones that can be explained by a phylogenetic tree are assigned non-negligible probabilities. The proposed Bayesian framework implies posterior distributions on the number of subclones, their genotypes, cellular proportions and the phylogenetic tree spanned by the inferred subclones. The proposed method is validated against different sets of simulated and real-world data using single and multiple tumor samples. An open source software package is available at <http://www.compgenome.org/treeclone>.

REFERENCES

- ADAMS, R. P., GHAMRANI, Z. and JORDAN, M. I. (2010). Tree-structured stick breaking for hierarchical data. In *Advances in Neural Information Processing Systems* 19–27.
- APARICIO, S. and CALDAS, C. (2013). The implications of clonal genome evolution for cancer medicine. *N. Engl. J. Med.* **368** 842–851.
- BAFNA, V., GUSFIELD, D., LANCIA, G. and YOOSEPH, S. (2003). Haplotyping as perfect phylogeny: A direct approach. *J. Comput. Biol.* **10** 323–340.
- BONAVIA, R., CAVENEE, W. K., FURNARI, F. B. et al. (2011). Heterogeneity maintenance in glioblastoma: A social network. *Cancer Res.* **71** 4055–4060.
- BROCKS, D., ASSENOV, Y., MINNER, S., BOGATYROVA, O., SIMON, R., KOOP, C., OAKES, C., ZUCKNICK, M., LIPKA, D. B., WEISCHENFELDT, J. et al. (2014). Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Rep.* **8** 798–806.
- BRODERICK, T., KULIS, B. and JORDAN, M. (2013). MAD-Bayes: MAP-based asymptotic derivations from Bayes. In *Proceedings of the 30th International Conference on Machine Learning* 226–234.

Key words and phrases. Latent feature model, mutation pair, NGS data, phylogenetic tree, subclone, tumor heterogeneity.

- CARTER, S. L., CIBULSKIS, K., HELMAN, E., MCKENNA, A., SHEN, H., ZACK, T., LAIRD, P. W., ONOFRIO, R. C., WINCKLER, W. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30** 413–421.
- CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93** 935–948.
- DAGUM, L. and MENON, R. (1998). OpenMP: An industry standard API for shared-memory programming. *IEEE Comput. Sci. Eng.* **5** 46–55.
- DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika* **85** 363–377. [MR1649118](#)
- DESHWAR, A. G., VEMBU, S., YUNG, C. K., JANG, G. H., STEIN, L. and MORRIS, Q. (2015). PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16** 35.
- FAN, X., ZHOU, W., CHONG, Z., NAKHLEH, L. and CHEN, K. (2014). Towards accurate characterization of clonal heterogeneity based on structural variation. *BMC Bioinform.* **15** 1.
- FISCHER, A., VÁZQUEZ-GARCÍA, I., ILLINGWORTH, C. J. R. and MUSTONEN, V. (2014). High-definition reconstruction of clonal composition in cancer. *Cell Rep.* **7** 1740–1752.
- GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* 457–472.
- GEYER, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface* 156–163. Interface Foundation of North America, Fairfax Station, VA.
- GIORDANO, R. J., BRODERICK, T. and JORDAN, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems* 1441–1449.
- JIAO, W., VEMBU, S., DESHWAR, A. G., STEIN, L. and MORRIS, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinform.* **15** 35.
- JOHNSON, V. E. (2004). A Bayesian χ^2 test for goodness-of-fit. *Ann. Statist.* **32** 2361–2384. [MR2153988](#)
- LEE, J., MÜLLER, P., GULUKOTA, K. and JI, Y. (2015). A Bayesian feature allocation model for tumor heterogeneity. *Ann. Appl. Stat.* **9** 621–639. [MR3371328](#)
- LI, H. and DURBIN, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25** 1754–1760.
- MARASS, F., MOULIERE, F., YUAN, K., ROSENFELD, N. and MARKOWETZ, F. (2016). A phylogenetic latent feature model for clonal deconvolution. *Ann. Appl. Stat.* **10** 2377–2404. [MR3592061](#)
- MARUSYK, A., ALMENDRO, V. and POLYAK, K. (2012). Intra-tumour heterogeneity: A looking glass for cancer? *Nat. Rev. Cancer* **12** 323–334.
- MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20** 1297–1303.
- MILLER, K. T., GRIFFITHS, T. L. and JORDAN, M. I. (2008). The phylogenetic Indian buffet process: A non-exchangeable nonparametric prior for latent features. In *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence* 403–410.
- MILLER, C. A., WHITE, B. S., DEES, N. D., GRIFFITH, M., WELCH, J. S., GRIFFITH, O. L., VIJ, R., TOMASSON, M. H., GRAUBERT, T. A., WALTER, M. J. et al. (2014). SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.* **10** e1003665.
- NIK-ZAINAL, S., VAN LOO, P., WEDGE, D. C., ALEXANDROV, L. B., GREENMAN, C. D., LAU, K. W., RAINE, K., JONES, D., MARSHALL, J., RAMAKRISHNA, M. et al. (2012). The life history of 21 breast cancers. *Cell* **149** 994–1007.

- NOWELL, P. C. (1976). The clonal evolution of tumor cell populations. *Science* **194** 23–28.
- O’HAGAN, A. (1995). Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. Ser. B* **57** 99–138. [MR1325379](#)
- ROTH, A., KHATTRA, J., YAP, D., WAN, A., LAKS, E., BIELE, J., HA, G., APARICIO, S., BOUCHARD-CÔTÉ, A. and SHAH, S. P. (2014). PyClone: Statistical inference of clonal population structure in cancer. *Nat. Methods* **11** 396–398.
- SCHWARZ, R. F., NG, C. K., COOKE, S. L., NEWMAN, S., TEMPLE, J., PISKORZ, A. M., GALE, D., SAYAL, K., MURTAZA, M., BALDWIN, P. J. et al. (2015). Spatial and temporal heterogeneity in high-grade serous ovarian cancer: A phylogenetic analysis. *PLoS Med.* **12** e1001789.
- SENGUPTA, S., WANG, J., LEE, J., MÜLLER, P., GULUKOTA, K., BANERJEE, A. and JI, Y. (2015). BayClone: Bayesian nonparametric inference of tumor subclones using NGS data. In *Proceedings of the Pacific Symposium on Biocomputing (PSB)* **20** 467–478.
- SENGUPTA, S., GULUKOTA, K., ZHU, Y., OBER, C., NAUGHTON, K., WENTWORTH-SHEILDS, W. and JI, Y. (2016). Ultra-fast local-haplotype variant calling using paired-end DNA-sequencing data reveals somatic mosaicism in tumor and normal blood samples. *Nucleic Acids Res.* **44** e25.
- VAN LOO, P., NORDGARD, S. H., LINGJÆRDE, O. C., RUSSNES, H. G., RYE, I. H., SUN, W., WEIGMAN, V. J., MARYNEN, P., ZETTERBERG, A., NAUME, B. et al. (2010). Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. USA* **107** 16910–16915.
- XU, Y., MÜLLER, P., YUAN, Y., GULUKOTA, K. and JI, Y. (2015). MAD Bayes for tumor heterogeneity—feature allocation with exponential family sampling. *J. Amer. Statist. Assoc.* **110** 503–514. [MR3367243](#)
- ZARE, H., WANG, J., HU, A., WEBER, K., SMITH, J., NICKERSON, D., SONG, C., WITTEN, D., BLAU, C. A. and NOBLE, W. S. (2014). Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.* **10** e1003703.
- ZHOU, T., MÜLLER, P., SENGUPTA, S. and JI, Y. (2019). PairClone: A Bayesian subclone caller based on mutation pairs. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 705–725.
- ZHOU, T., SENGUPTA, S., MÜLLER, P. and JI, Y. (2019). Supplement to “TreeClone: Reconstruction of Tumor Subclone Phylogeny Based on Mutation Pairs using Next Generation Sequencing Data.” DOI:[10.1214/18-AOAS1224SUPP](#).

LATENT SPACE MODELLING OF MULTIDIMENSIONAL NETWORKS WITH APPLICATION TO THE EXCHANGE OF VOTES IN EUROVISION SONG CONTEST

BY SILVIA D'ANGELO*, THOMAS BRENDAN MURPHY[†] AND MARCO ALFÒ*

Sapienza, University of Rome and University College Dublin[†]*

The Eurovision Song Contest is a popular TV singing competition held annually among country members of the European Broadcasting Union. In this competition, each member can be both contestant and jury, as it can participate with a song and/or vote for other countries' tunes. During the years, the voting system has repeatedly been accused of being biased by tactical voting; votes would represent strategic interests rather than actual musical preferences of the voting countries. In this work, we develop a latent space model to investigate the presence of a latent structure underlying the exchange of votes. Focusing on the period from 1998 to 2015, we represent the vote exchange as a multivariate network: each edition is a network, where countries are the nodes and two countries are linked by an edge if one voted for the other. The different networks are taken to be independent replicates of a conditional Bernoulli distribution, with success probability specified as a function of a common latent space capturing the overall relationships among the countries. Proximity denotes similarity, and countries close in the latent space are more likely to exchange votes. If the exchange of votes depends on the similarity between countries, the quality of the competing songs might not be a relevant factor in the determination of the voting preferences, and this would suggest the presence of some bias. A Bayesian hierarchical modelling approach is employed to estimate the parameters, where the probability of a connection between any two countries is a function of their distance in the latent space, network-specific parameters and edge-specific covariates. The estimated latent space is found to be relevant in the determination of edge probabilities, however, the positions of the countries in such space only partially correspond to their actual geographical positions.

REFERENCES

- AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed-membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.
- BLANGIARDO, M. and BAIIO, G. (2014). Evidence of bias in the Eurovision song contest: Modelling the votes using Bayesian hierarchical models. *J. Appl. Stat.* **41** 2312–2322. [MR3292674](#)
- BUTTS, C. T. and CARLEY, K. M. (2005). Some simple algorithms for structural comparison. *Computational and Mathematical Organization Theory* **11** 291–305.
- CLERIDES, S. and STENGOS, T. (2006). Love thy neighbour, love thy kin: Strategy and bias in the Eurovision song contest. *Discussion Paper Series of Centre for Economic Policy Research* **5732** 1–28.

Key words and phrases. Eurovision, latent space models, multidimensional networks.

- D'ANGELO, S., MURPHY, T. B. and ALFÒ, M. (2019). Supplement to "Latent space modelling of multidimensional networks with application to the exchange of votes in Eurovision song contest." DOI:[10.1214/18-AOAS1221SUPP](https://doi.org/10.1214/18-AOAS1221SUPP).
- DRYDEN, I. L. and MARDIA, K. V. (1998). *Statistical Shape Analysis*. *Wiley Series in Probability and Statistics: Probability and Statistics*. Wiley, Chichester. [MR1646114](#)
- DURANTE, D., DUNSON, D. B. and VOGELSTEIN, J. T. (2017). Nonparametric Bayes modeling of populations of networks. *J. Amer. Statist. Assoc.* **112** 1516–1530. [MR3750873](#)
- ERDŐS, P. and RÉNYI, A. (1959). On random graphs. I. *Publ. Math. Debrecen* **6** 290–297. [MR0120167](#)
- ERDŐS, P. and RÉNYI, A. (1960). On the evolution of random graphs. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* **5** 17–61. [MR0125031](#)
- FENN, D., SULEMANA, O., EFSTATHIOUB, J. and JOHNSON, N. F. (2006). How does Europe make its mind up? Connections, cliques, and compatibility between countries in the Eurovision song contest. *Phys. A* **360** 576–598.
- FIENBERG, S. E., MEYER, M. M. and WASSERMAN, S. S. (1985). Statistical analysis of multiple sociometric relations. *J. Amer. Statist. Assoc.* **80** 51–67.
- FRANK, O. and STRAUSS, D. (1986). Markov graphs. *J. Amer. Statist. Assoc.* **81** 832–842. [MR0860518](#)
- GINSBURGH, V. and NOURY, A. (2008). The Eurovision song contest. Is voting political or cultural?. *European Journal of Political Economy* **24** 41–52.
- GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.
- GOLLINI, I. and MURPHY, T. B. (2016). Joint modeling of multiple network views. *J. Comput. Graph. Statist.* **25** 246–265. [MR3474046](#)
- GREENE, D. and CUNNINGHAM, P. (2013). Producing a unified graph representation from multiple social network views. *Proceedings of the 5th Annual ACM Web Science Conference (Web-Sci'13)* 118–121.
- HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354. [MR2364300](#)
- HOFF, P. D. (2005). Bilinear mixed-effects models for dyadic data. *J. Amer. Statist. Assoc.* **100** 286–295. [MR2156838](#)
- HOFF, P. D. (2011). Hierarchical multilinear models for multiway data. *Comput. Statist. Data Anal.* **55** 530–543. [MR2736574](#)
- HOFF, P. (2015). Dyadic data analysis with amen.
- 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](#)
- HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76** 33–65. [MR0608176](#)
- HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. [MR0718088](#)
- KRIVITSKY, P. N., HANDCOCK, M. S., RAFTERY, A. E. and HOFF, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Soc. Netw.* **31** 204–213.
- LYNCH, K. (2015). Eurovision recognised by Guinness world records as the longest-running annual TV music competition (international). Available at <http://www.guinnessworldrecords.com/news/2015/5/eurovision-recognised-by-guinness-world-records-as-the-longest-running-annual-tv-379520>.
- MANTZARIS, A. V., REIN, S. R. and HOPKINS, A. D. (2018). Examining collusion and voting biases between countries during the Eurovision song contest since 1957. *Journal of Artificial Societies and Social Simulation* **21**.

- MURPHY, T. B. (2016). Model-based clustering for network data. In *Handbook of Cluster Analysis*. Chapman & Hall/CRC Handb. Mod. Stat. Methods 337–357. CRC Press, Boca Raton, FL. [MR3644719](#)
- ROBINS, G., PATTISON, P., KALISH, Y. and LUSHER, D. (2007). An introduction to exponential random graph (p^*) models for social network. *Soc. Netw.* **29** 173–191.
- ROSÉN, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement. I. *Ann. Math. Stat.* **4** 373–397. [MR0321223](#)
- SAAVEDRAA, S., EFSTATHIOU, J. and REED-TSOCHAS, F. (2007). Identifying the underlying structure and dynamic interactions in a voting network. *Phys. A* **377** 672–688.
- SALTER-TOWNSHEND, M. and MCCORMICK, T. H. (2017). Latent space models for multiview network data. *Ann. Appl. Stat.* **11** 1217–1244. [MR3709558](#)
- SALTER-TOWNSHEND, M., WHITE, A., GOLLINI, I. and MURPHY, T. B. (2012). Review of statistical network analysis: Models, algorithms, and software. *Stat. Anal. Data Min.* **5** 260–264. [MR2958152](#)
- SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14** 75–100. [MR1449742](#)
- SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. [MR1979380](#)
- SPIERDIJK, L. and VELLEKOOP, M. (2006). Geography, culture, and religion: Explaining the bias in Eurovision song contest voting. *Memorandum, Department of Applied Mathematics, University of Twente, Enschede* **1794** 1–30.
- SWEET, T. M., THOMAS, A. C. and JUNKER, B. W. (2013). Hierarchical network models for education research: Hierarchical latent space models. *Journal of Educational and Behavioural Statistics* **38** 295–318.
- YAIR, G. (1995). “Unite unite Europe” the political and cultural structures of Europe as reflected in the Eurovision song contest. *Soc. Netw.* **17** 147–161.

MODELING ASSOCIATION IN MICROBIAL COMMUNITIES WITH CLIQUE LOGLINEAR MODELS¹

BY ADRIAN DOBRA^{*,2}, CAMILO VALDES^{†,2}, DRAGANA AJDIC[‡],
 BERTRAND CLARKE[§] AND JENNIFER CLARKE[§]

University of Washington^{}, Florida International University[†], University of
 Miami[‡] and University of Nebraska-Lincoln[§]*

There is a growing awareness of the important roles that microbial communities play in complex biological processes. Modern investigation of these often uses next generation sequencing of metagenomic samples to determine community composition. We propose a statistical technique based on clique loglinear models and Bayes model averaging to identify microbial components in a metagenomic sample at various taxonomic levels that have significant associations. We describe the model class, a stochastic search technique for model selection, and the calculation of estimates of posterior probabilities of interest. We demonstrate our approach using data from the Human Microbiome Project and from a study of the skin microbiome in chronic wound healing. Our technique also identifies significant dependencies among microbial components as evidence of possible microbial syntrophy.

REFERENCES

- ABRAMOWITZ, M. and STEGUN, I. A. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. U.S. Dept. of Commerce: US GPO, Washington, DC.
- BARRY, D. and HARTIGAN, J. A. (1992). Product partition models for change point problems. *Ann. Statist.* **20** 260–279. [MR1150343](#)
- BERGER, J. O., GHOSH, J. K. and MUKHOPADHYAY, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Plann. Inference* **112** 241–258. [MR1961733](#)
- BHATTACHARYA, A. and DUNSON, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *J. Amer. Statist. Assoc.* **107** 362–377. [MR2949366](#)
- BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (2007). *Discrete Multivariate Analysis: Theory and Practice*. Springer, New York. [MR2344876](#)
- CANALE, A. and DUNSON, D. B. (2011). Bayesian kernel mixtures for counts. *J. Amer. Statist. Assoc.* **106** 1528–1539. [MR2896854](#)
- CARVALHO, C. M. and SCOTT, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96** 497–512. [MR2538753](#)
- CHARUVAKA, A. and RANGWALA, H. (2011). Evaluation of short read metagenomic assembly. *BMC Genomics* **12** S8.
- CLARKE, B., VALDES, C., DOBRA, A. and CLARKE, J. (2015). A Bayes testing approach to metagenomic profiling in bacteria. *Stat. Interface* **8** 173–185. [MR3322164](#)
- DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. [MR1241267](#)

Key words and phrases. Contingency tables, graphical models, model selection, microbiome, next generation sequencing.

- DE KIEVIT, T. and IGLEWSKI, B. (2000). Bacterial quorum sensing in pathogenic relationships. *Infect. Immun.* **68** 4839–4849.
- DELLAPORTAS, P. and FORSTER, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86** 615–633. [MR1723782](#)
- DELLAPORTAS, P. and TARANTOLA, C. (2005). Model determination for categorical data with factor level merging. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 269–283. [MR2137325](#)
- DOBRA, A. (2009). Variable selection and dependency networks for genomewide data. *Biostatistics* **10** 621–639.
- DOBRA, A. and LENKOSKI, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.* **5** 969–993. [MR2840183](#)
- DOBRA, A. and MASSAM, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. *Stat. Methodol.* **7** 240–253. [MR2643600](#)
- DOBRA, A., VALDES, C., AJDIC, D., CLARKE, B. and CLARKE, J. (2019). Supplement to “Modeling association in microbial communities with clique loglinear models.” DOI:[10.1214/18-AOAS1229SUPP](#).
- DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Amer. Statist. Assoc.* **104** 1042–1051. [MR2562004](#)
- EDWARDS, D. and HAVRÁNEK, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika* **72** 339–351. [MR0801773](#)
- FETTWEIS, J., SERRANO, M., GIRERD, P., JEFFERSON, K. and BUCK, G. (2012). A new era of the vaginal microbiome: Advances using next generation sequencing. *Chem. Biodivers.* **9** 965–976.
- FIENBERG, S. E. and RINALDO, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *J. Statist. Plann. Inference* **137** 3430–3445. [MR2363267](#)
- FIERER, N., LAUBER, C., ZHOU, N., McDONALD, D., COSTELLO, E. and KNIGHT, R. (2010). Forensic identification using skin bacterial communities. *Proc. Natl. Acad. Sci. USA* **107** 6477–6481.
- NIH HMP WORKING GROUP, PETERSON, J., GARGES, S., GIOVANNI, M., MCINNES, P., WANG, L., SCHLOSS, J. A., BONAZZI, V., MCEWEN, J. E. et al. (2009). The NIH human microbiome project. *Genome Res.* **19** 2317–2323.
- HANKIN, R. K. S. (2006). Additive integer partitions in R. *J. Stat. Softw.* **16**. Code Snippet 1.
- HANS, C., DOBRA, A. and WEST, M. (2007). Shotgun stochastic search for “large p ” regression. *J. Amer. Statist. Assoc.* **102** 507–516. [MR2370849](#)
- HASMAN, H., SAPUTRA, D., SICHERITZ-PONTEN, T., LUND, O., SVENDSEN, C. A., FRIMODT-MØLLER, N. and AARESTRUP, F. M. (2014). Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *Eur. J. Clin. Microbiol. Infect. Dis.* **52** 139–146.
- HOFFMANN, C., DOLLIVE, S., GRUNBERG, S., CHEN, J., LI, H., WU, G., LEWIS, J. and BUSHMAN, F. (2013). Archaea and fungi of the human gut microbiome: Correlations with diet and bacterial residents. *PLoS ONE* **8** e66019.
- HUANG, B., FETTWEIS, J., BROOKS, J. P., JEFFERSON, K. and BUCK, G. (2014). The changing landscape of the vaginal microbiome. *Clin. Lab. Med.* **34** 747–761.
- JOHNDROW, J. E., BHATTACHARYA, A. and DUNSON, D. B. (2017). Tensor decompositions and sparse log-linear models. *Ann. Statist.* **45** 1–38. [MR3611485](#)
- JONES, B., CARVALHO, C., DOBRA, A., HANS, C., CARTER, C. and WEST, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.* **20** 388–400. [MR2210226](#)
- KOCH, G., NADAL-JIMENEZ, P., REIS, C., MUNTENDAM, R., BOKHOVE, M., MELILLO, E., DIJKSTRA, B., COOL, R. and QUAX, W. (2014). Reducing virulence of the human pathogen *Burkholderia* by altering the substrate specificity of the quorum-quenching acylase PvdQ. *Proc. Natl. Acad. Sci. USA* **111** 1568–1573.

- KUNIHAMA, T. and DUNSON, D. B. (2013). Bayesian modeling of temporal dependence in large sparse contingency tables. *J. Amer. Statist. Assoc.* **108** 1324–1338. [MR3174711](#)
- LANGMEAD, B. and SALZBERG, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9** 357–359.
- LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. Clarendon Press, Oxford. [MR1419991](#)
- LENKOSKI, A. and DOBRA, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *J. Comput. Graph. Statist.* **20** 140–157. [MR2816542](#)
- LETAC, G. and MASSAM, H. (2012). Bayes factors and the geometry of discrete hierarchical log-linear models. *Ann. Statist.* **40** 861–890. [MR2985936](#)
- LEVY, R. and BORENSTEIN, E. (2013). Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc. Natl. Acad. Sci. USA* **110** 12804–12809.
- LOVATO, P. (2015). Bag of words approaches for Bioinformatics Ph. D. thesis, Dept. Informatics, Univ. Verona.
- MADIGAN, D., GAVRIN, J. and RAFTERY, A. E. (1995). Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Comm. Statist. Theory Methods* **24** 2271–2292. [MR1350662](#)
- MADIGAN, D. and RAFTERY, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occam’s window. *J. Amer. Statist. Assoc.* **89** 1535–1546.
- MADIGAN, D. and YORK, J. (1995). Bayesian graphical models for discrete data. *Int. Stat. Rev.* **63** 215–232.
- MADIGAN, D. and YORK, J. C. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika* **84** 19–31. [MR1450189](#)
- MARKOWITZ, V. M., CHEN, I. M., PALANIAPPAN, K., CHU, K., SZETO, E., PILLAY, M., RATNER, A., HUANG, J., WOYKE, T. et al. (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.* **42** D560–D567.
- MASSAM, H., LIU, J. and DOBRA, A. (2009). A conjugate prior for discrete hierarchical log-linear models. *Ann. Statist.* **37** 3431–3467. [MR2549565](#)
- MINOT, S., BRYSON, A., CHEHOUD, C., WU, G., LEWIS, J. and BUSHMAN, F. (2013). Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. USA* **110** 12450–12455.
- MOHAMMADI, A., ABEGAZ, F., VAN DEN HEUVEL, E. and WIT, E. C. (2017). Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 629–645. [MR3632345](#)
- MUKHERJEE, C. and RODRIGUEZ, A. (2016). GPU-powered shotgun stochastic search for Dirichlet process mixtures of Gaussian graphical models. *J. Comput. Graph. Statist.* **25** 762–788. [MR3533637](#)
- NGUYEN, N.-P., WARNOW, T., POP, M. and WHITE, B. (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms and Microbiomes* **2** 16004.
- ONORANTE, L. and RAFTERY, A. E. (2016). Dynamic model averaging in large model spaces using dynamic Occam’s window. *Eur. Econ. Rev.* **81** 2–14.
- RANJAN, R., RANI, A., METWALLY, A., MCGEE, H. S. and PERKINS, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* **469** 967–977.
- ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, D. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. [MR0468014](#)
- SMITH, K., COLLIER, A., TOWNSEND, E. M., O’DONNELL, L. E., BAL, A. M., BUTCHER, J., MACKAY, W. G., RAMAGE, G. and WILLIAMS, C. (2016). One step closer to understanding the

- role of bacteria in diabetic foot ulcers: Characterising the microbiome of ulcers. *BMC Microbiol.* **16** 54.
- TARANTOLA, C. (2004). MCMC model determination for discrete graphical models. *Stat. Model.* **4** 39–61. [MR2037813](#)
- THOENDEL, M., JERALDO, P. R., GREENWOOD-QUAINTANCE, K. E., YAO, J. Z., CHIA, N., HANSEN, A. D., ABDEL, M. P. and PATEL, R. (2016). Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. *J. Microbiol. Methods* **127** 141–145.
- WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. *Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. Wiley, Chichester. [MR1112133](#)
- ZHAO, J., SCHLOSS, P., KALIKIN, L., CARMODY, L., FOSTER, B., PETROSINO, J., CAVALCOLI, J., VANDEVANTER, D., MURRAY, S. et al. (2012). Decade-long bacterial community dynamics in cystic fibrosis airway. *Proc. Natl. Acad. Sci. USA* **109** 5809–5814.
- 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](#)
- ZHOU, J., HERRING, A. H., BHATTACHARYA, A., OLSHAN, A. F., DUNSON, D. B. and THE NATIONAL BIRTH DEFECTS PREVENTION STUDY (2016). Nonparametric Bayes modeling for case control studies with many predictors. *Biometrics* **72** 184–192. [MR3500587](#)

VARIABLE PRIORITIZATION IN NONLINEAR BLACK BOX METHODS: A GENETIC ASSOCIATION CASE STUDY¹

BY LORIN CRAWFORD^{*}, SETH R. FLAXMAN[†], DANIEL E. RUNCIE[‡] AND
MIKE WEST[§]

Brown University^{}, Imperial College London[†], University of California, Davis[‡]
and Duke University[§]*

The central aim in this paper is to address variable selection questions in nonlinear and nonparametric regression. Motivated by statistical genetics, where nonlinear interactions are of particular interest, we introduce a novel and interpretable way to summarize the relative importance of predictor variables. Methodologically, we develop the “RelATive cEntrality” (RATE) measure to prioritize candidate genetic variants that are not just marginally important, but whose associations also stem from significant covarying relationships with other variants in the data. We illustrate RATE through Bayesian Gaussian process regression, but the methodological innovations apply to other “black box” methods. It is known that nonlinear models often exhibit greater predictive accuracy than linear models, particularly for phenotypes generated by complex genetic architectures. With detailed simulations and two real data association mapping studies, we show that applying RATE enables an explanation for this improved performance.

REFERENCES

- ALAA, A. M. and VAN DER SCHAAR, M. (2017). Bayesian nonparametric causal inference: Information rates and learning algorithms. Available at [ArXiv:1712.08914](https://arxiv.org/abs/1712.08914).
- ANKRA-BADU, G. A., POMP, D., SHRINER, D., ALLISON, D. B. and YI, N. (2009). Genetic influences on growth and body composition in mice: Multilocus interactions. *Int. J. Obes.* **33** 89–95. DOI:[10.1038/ijo.2008.215](https://doi.org/10.1038/ijo.2008.215).
- BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. [MR2065192](https://arxiv.org/abs/2006.0192)
- BROCKMANN, G. A., HALEY, C. S., RENNE, U., KNOTT, S. A. and SCHWERIN, M. (1998). Quantitative trait loci affecting body weight and fatness from a mouse line selected for extreme high growth. *Genetics* **150** 369–381.
- BROSS, C. D., HOWES, T. R., ABOLHASSANI RAD, S., KLJAKIC, O. and KOHALMI, S. E. (2017). Subcellular localization of *Arabidopsis* argonate dehydratases suggests novel and non-enzymatic roles. *J. Exp. Bot.* **68** 1425–1440.
- CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. [MR2650751](https://arxiv.org/abs/2006.0751)
- CARVALHO, C. M. and WEST, M. (2007). Dynamic matrix-variate graphical models. *Bayesian Anal.* **2** 69–97. [MR2289924](https://arxiv.org/abs/2007.28924)

Key words and phrases. Nonlinear regression, Gaussian processes, centrality measures, variable prioritization, genome-wide association studies, statistical genetics.

- CHAUDHURI, A., KAKDE, D., SADEK, C., GONZALEZ, L. and KONG, S. (2017). The mean and median criterion for automatic kernel bandwidth selection for support vector data description. Available at [arXiv:1708.05106](https://arxiv.org/abs/1708.05106).
- CHEN, X., MCCLUSKY, R., CHEN, J., BEAVEN, S. W., TONTONOV, P., ARNOLD, A. P. and REUE, K. (2012). The number of X chromosomes causes sex differences in adiposity in mice. *PLoS Genet.* **8** e1002709.
- CHEN, X., MCCLUSKY, R., ITOH, Y., REUE, K. and ARNOLD, A. P. (2013). X and Y chromosome complement influence adiposity and metabolism in mice. *Endocrinology* **154** 1092–1104. DOI:[10.1210/en.2012-2098](https://doi.org/10.1210/en.2012-2098).
- 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/10.1214/1099.4603)
- COTTER, A., KESHET, J. and SREBRO, N. (2011). Explicit approximations of the Gaussian kernel. Available at [arXiv:1109.4603](https://arxiv.org/abs/1109.4603).
- COX, K. H., BONTJUIS, P. J. and RISSMAN, E. F. (2014). Mouse model systems to study sex chromosome genes and behavior: Relevance to humans. *Front. Neuroendocrinol.* **35** 405–419. DOI:[10.1016/j.yfrne.2013.12.004](https://doi.org/10.1016/j.yfrne.2013.12.004).
- CRAWFORD, L. and ZHOU, X. (2018). Genome-wide marginal epistatic association mapping in case-control studies. *BioRxiv* 374983.
- CRAWFORD, L., ZENG, P., MUKHERJEE, S. and ZHOU, X. (2017). Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLoS Genet.* **13** e1006869.
- CRAWFORD, L., WOOD, K. C., ZHOU, X. and MUKHERJEE, S. (2018). Bayesian approximate kernel regression with variable selection. *J. Amer. Statist. Assoc.* **113** 1710–1721. [MR3902240](https://doi.org/10.1080/01621459.2018.1500000)
- CRAWFORD, L., FLAXMAN, S. R., RUNCIE, D. E. and WEST, M. (2019). Supplement to “Variable prioritization in nonlinear black box methods: A genetic association case study.” DOI:[10.1214/18-AOAS1222SUPPA](https://doi.org/10.1214/18-AOAS1222SUPPA), DOI:[10.1214/18-AOAS1222SUPPB](https://doi.org/10.1214/18-AOAS1222SUPPB), DOI:[10.1214/18-AOAS1222SUPPC](https://doi.org/10.1214/18-AOAS1222SUPPC), DOI:[10.1214/18-AOAS1222SUPPD](https://doi.org/10.1214/18-AOAS1222SUPPD).
- CUEVAS, J., CROSSA, J., MONTESINOS-LÓPEZ, O. A., BURGUEÑO, J., PÉREZ-RODRÍGUEZ, P. and DE LOS CAMPOS, G. (2017). Bayesian genomic prediction with genotype × environment interaction kernel models. *G3 (Bethesda)* **7** 41–53.
- DEMETRASHVILI, N., DEN HEUVEL, E. R. V. and WIT, E. C. (2013). Probability genotype imputation method and integrated weighted lasso for QTL identification. *BMC Genet.* **14** 125.
- DE LOS CAMPOS, G., NAYA, H., GIANOLA, D., CROSSA, J., LEGARRA, A., MANFREDI, E., WEIGEL, K. and COTES, J. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* **182** 375–385.
- DE LOS CAMPOS, G., GIANOLA, D., ROSA, G. J. M., WEIGEL, K. A. and CROSSA, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet. Res.* **92** 295–308.
- DIAMENT, A. L. and WARDEN, C. H. (2003). Multiple linked mouse chromosome 7 loci influence body fat mass. *Int. J. Obes.* **28** 199 EP.
- DRINEAS, P. and MAHONEY, M. W. (2005). On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.* **6** 2153–2175. [MR2249884](https://doi.org/10.1162/0899766052849884)
- FASSHAUER, G. and MCCOURT, M. (2016). *Kernel-Based Approximation Methods Using MATLAB*. World Scientific, Hackensack, NJ.
- GELMAN, A., HWANG, J. and VEHTARI, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24** 997–1016. [MR3253850](https://doi.org/10.1007/s11222-013-9375-0)
- GOUTIS, C. and ROBERT, C. P. (1998). Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections. *Biometrika* **85** 29–37. [MR1627250](https://doi.org/10.1017/S0007122600005200)
- 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-BA1004)

- GRUBER, L. F. and WEST, M. (2017). Bayesian online variable selection and scalable multivariate volatility forecasting in simultaneous graphical dynamic linear models. *Econ. Stat.* **3** 3–22. [MR3666239](#)
- GUAN, Y. and STEPHENS, M. (2011). Bayesian variable selection regression for Genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5** 1780–1815. [MR2884922](#)
- HEMANI, G., KNOTT, S. and HALEY, C. (2013). An evolutionary perspective on epistasis and the missing heritability. *PLoS Genet.* **9** e1003295.
- HEMANI, G., SHAKHBAZOV, K., WESTRA, H.-J., ESKO, T., HENDERS, A. K., MCRAE, A. F., YANG, J., GIBSON, G., MARTIN, N. G., METSPALU, A., FRANKE, L., MONTGOMERY, G. W., VISSCHER, P. M. and POWELL, J. E. (2014). Detection and replication of epistasis influencing transcription in humans. *Nature* **508** 249–253.
- HILL, W. G., GODDARD, M. E. and VISSCHER, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4** e1000008.
- HORN, T., SANDMANN, T., FISCHER, B., AXELSSON, E., HUBER, W. and BOUTROS, M. (2011). Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. *Nat. Methods* **8** 341–346.
- HOU, Q. and BARTELS, D. (2015). Comparative study of the aldehyde dehydrogenase (ALDH) gene superfamily in the glycophyte *Arabidopsis thaliana* and *Eutrema halophytes*. *Ann. Bot.* **115** 465–479.
- HOWARD, R., CARRIQUIRY, A. L. and BEAVIS, W. D. (2014). Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3 (Bethesda)* **4** 1027–1046.
- JIANG, Y. and REIF, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* **201** 759–768.
- KANG, H. M., SUL, J. H., SERVICE, S. K., ZAITLEN, N. A., KONG, S.-Y., FREIMER, N. B., SABATTI, C. and ESKIN, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42** 348–354.
- KIM, S. V., MEHAL, W. Z., DONG, X., HEINRICH, V., PYPAERT, M., MELLMAN, I., DEMBO, M., MOOSEKER, M. S., WU, D. and FLAVELL, R. A. (2006). Modulation of cell adhesion and motility in the immune system by Myo1f. *Science* **314** 136–139.
- KIRCH, H.-H., BARTELS, D., WEI, Y., SCHNABLE, P. S. and WOOD, A. J. (2004). The ALDH gene superfamily of Arabidopsis. *Trends Plant Sci.* **9** 371–377.
- KLEYN, P. W., FAN, W., KOVATS, S. G., LEE, J. J., PULIDO, J. C., WU, Y., BERKEMEIER, L. R., MISUMI, D. J., HOLMGREN, L. et al. (1996). Identification and characterization of the mouse obesity gene *tubby*: A member of a novel gene family. *Cell* **85** 281–290.
- KOLMOGOROV, A. N. and ROZANOV, JU. A. (1960). On a strong mixing condition for stationary Gaussian processes. *Theory Probab. Appl.* **5** 222–227. [MR0133175](#)
- LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. [MR2420243](#)
- LIM, C. and YU, B. (2016). Estimation stability with cross-validation (ESCV). *J. Comput. Graph. Statist.* **25** 464–492. [MR3499690](#)
- LIN, L., CHAN, C. and WEST, M. (2016). Discriminative variable subsets in Bayesian classification with mixture models, with application in flow cytometry studies. *Biostatistics* **17** 40–53. [MR3449849](#)
- LIPPERT, C., LISTGARTEN, J., LIU, Y., KADIE, C. M., DAVIDSON, R. I. and HECKERMAN, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8** 833–835.
- LOUDET, O., CHAILLOU, S., CAMILLERI, C., BOUCHEZ, D. and DANIEL-VEDELE, F. (2002). Bay-0 $\times$ Shahdara recombinant inbred line population: A powerful tool for the genetic dissection of complex traits in *Arabidopsis*. *Theor. Appl. Genet.* **104** 1173–1184.
- MACKAY, T. F. C. (2014). Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. *Nat. Rev. Genet.* **15** 22–33.

- MATHAI, A. M. and PROVOST, S. B. (1992). *Quadratic Forms in Random Variables. Theory and Applications. Statistics: Textbooks and Monographs* **126**. Dekker, New York. [MR1192786](#)
- MERCER, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. R. Soc. Lond. Ser. A* **209** 415–446.
- PAIGEN, B., MITCHELL, D., REUE, K., MORROW, A., LUSIS, A. J. and LEBOEUF, R. C. (1987). Ath-1, a gene determining atherosclerosis susceptibility and high density lipoprotein levels in mice. *Proc. Natl. Acad. Sci. USA* **84** 3763–3767.
- PHILLIPS, P. C. (2008). Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9** 855–867. DOI:[10.1038/nrg2452](#).
- PIIRONEN, J. and VEHTARI, A. (2016). Projection predictive model selection for Gaussian processes. In *IEEE International Workshop on Machine Learning for Signal Processing* 1–6. IEEE, New York.
- PIIRONEN, J. and VEHTARI, A. (2017). Comparison of Bayesian predictive methods for model selection. *Stat. Comput.* **27** 711–735. [MR3613594](#)
- PILLAI, N. S., WU, Q., LIANG, F., MUKHERJEE, S. and WOLPERT, R. L. (2007). Characterizing the function space for Bayesian kernel models. *J. Mach. Learn. Res.* **8** 1769–1797. [MR2332448](#)
- PRABHU, S. and PE’ER, I. (2012). Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome Res.* **22** 2230–2240.
- PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. W., DALY, M. J. and SHAM, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81** 559–575. DOI:[10.1086/519795](#).
- RAHIMI, A. and RECHT, B. (2007). Random features for large-scale kernel machines. *Adv. Neural Inf. Process. Syst.* **3** 5.
- RANCE, K. A., HILL, W. G. and KEIGHTLEY, P. D. (1997). Mapping quantitative trait loci for body weight on the X chromosome in mice. I. Analysis of a reciprocal F2 population. *Genet. Res.* **70** 117–124.
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR2514435](#)
- RICHARD, M. D. and LIPPMANN, R. P. (1991). Neural network classifiers estimate Bayesian *a posteriori* probabilities. *Neural Comput.* **3** 461–483.
- SCHÖLKOPF, B., HERBRICH, R. and SMOLA, A. J. (2001). A generalized representer theorem. In *Computational Learning Theory (Amsterdam, 2001). Lecture Notes in Computer Science* **2111** 416–426. Springer, Berlin. [MR2042050](#)
- SHI, J. Q., WANG, B., WILL, E. J. and WEST, R. M. (2012). Mixed-effects Gaussian process functional regression models with application to dose-response curve prediction. *Stat. Med.* **31** 3165–3177. [MR2993619](#)
- SMITH, A., NAIK, P. A. and TSAI, C.-L. (2006). Markov-switching model selection using Kullback–Leibler divergence. *J. Econometrics* **134** 553–577. [MR2328419](#)
- STEPHENS, M. and BALDING, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10** 681–690.
- SUDLOW, C., GALLACHER, J., ALLEN, N., BERAL, V., BURTON, P., DANESH, J., DOWNEY, P., ELLIOTT, P., GREEN, J. et al. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12** e1001779.
- TAN, S., CARUANA, R., HOOKER, G. and LOU, Y. (2017). Detecting bias in black-box models using transparent model distillation. Available at [arXiv:1710.06169](#).
- THE 1000 GENOMES PROJECT CONSORTIUM (2010). A map of human genome variation from population-scale sequencing. *Nature* **467** 1061–1073.
- THE WELLCOME TRUST CASE CONTROL CONSORTIUM (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447** 661–678.

- VALDAR, W., SOLBERG, L. C., GAUGUIER, D., BURNETT, S., KLENERMAN, P., COOKSON, W. O., TAYLOR, M. S., RAWLINS, J. N. P., MOTT, R. and FLINT, J. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.* **38** 879–887.
- WAHBA, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics **59**. SIAM, Philadelphia, PA. [MR1045442](#)
- WALDMANN, P., MÉSZÁROS, G., GREDLER, B., FÜRST, C. and SÖLKNER, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Front. Genet.* **4** 270.
- WAN, X., YANG, C., YANG, Q., XUE, H., FAN, X., TANG, N. L. and YU, W. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.* **87** 325–340.
- WANG, X., ELSTON, R. C. and ZHU, X. (2011a). Statistical interaction in human genetics: How should we model it if we are looking for biological interaction? *Nat. Rev. Genet.* **12** 74.
- WANG, X., ELSTON, R. C. and ZHU, X. (2011b). The meaning of interaction. *Hum. Hered.* **70** 269–277.
- WEISSBROD, O., GEIGER, D. and ROSSET, S. (2016). Multikernel linear mixed models for complex phenotype prediction. *Genome Res.* **26** 969–979.
- WENTZELL, A. M., ROWE, H. C., HANSEN, B. G., TICCONI, C., HALKIER, B. A. and KLIEBENSTEIN, D. J. (2007). Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *PLoS Genet.* **3** e162.
- WOO, J. H., SHIMONI, Y., YANG, W. S., SUBRAMANIAM, P., IYER, A., NICOLETTI, P., RODRÍGUEZ MARTÍNEZ, M., LÓPEZ, G., MATTIOLI, M. et al. (2015). Elucidating compound mechanism of action by network perturbation analysis. *Cell* **162** 441–451.
- WOOD, A. R., TUKE, M. A., NALLS, M. A., HERNANDEZ, D. G., BANDINELLI, S., SINGLETON, A. B., MELZER, D., FERRUCCI, L., FRAYLING, T. M. and WEEDON, M. N. (2014). Another explanation for apparent epistasis. *Nature* **514** E3–E5.
- 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, J., ZHAO, Q., YANG, Q., LIU, H., LI, Q., YI, X., CHENG, Y., GUO, L., FAN, C. and ZHOU, Y. (2016). Comparative transcriptomic analysis uncovers the complex genetic network for resistance to *Sclerotinia sclerotiorum* in *Brassica napus*. *Sci. Rep.* **6** 19007 EP.
- YALCIN, B., NICOD, J., BHOMRA, A., DAVIDSON, S., CLEAK, J., FARINELLI, L., ØSTERÅS, M., WHITLEY, A., YUAN, W. et al. (2010). Commercially available outbred mice for genome-wide association studies. *PLoS Genet.* **6** e1001085.
- YANDELL, B. S., MEHTA, T., BANERJEE, S., SHRINER, D., VENKATARAMAN, R., MOON, J. Y., NEELY, W. W., WU, H., VON SMITH, R. and YI, N. (2007). R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. *Bioinformatics* **23** 641–643. DOI:[10.1093/bioinformatics/btm011](#).
- YANG, J., ZAITLEN, N. A., GODDARD, M. E., VISSCHER, P. M. and PRICE, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46** 100–106.
- ZENG, P. and ZHOU, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nat. Commun.* **8** 456.
- ZHANG, Z., DAI, G. and JORDAN, M. I. (2011). Bayesian generalized kernel mixed models. *J. Mach. Learn. Res.* **12** 111–139. [MR2773550](#)
- ZHANG, Y. and LIU, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.* **39** 1167–1173.
- ZHANG, X., HUANG, S., ZOU, F. and WANG, W. (2010). TEAM: Efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics* **26** i217–i227. DOI:[10.1093/bioinformatics/btq186](#).
- ZHOU, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. *Ann. Appl. Stat.* **11** 2027–2051. [MR3743287](#)

ZHOU, X. and STEPHENS, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44** 821–825.

ZHOU, X. and STEPHENS, M. (2014). Efficient multivariate linear mixed model algorithms for genomewide association studies. *Nat. Methods* **11** 407–409.

COHERENCE-BASED TIME SERIES CLUSTERING FOR STATISTICAL INFERENCE AND VISUALIZATION OF BRAIN CONNECTIVITY

BY CAROLINA EUÁN, YING SUN AND HERNANDO OMBAO

King Abdullah University of Science and Technology (KAUST)

We develop the hierarchical cluster coherence (HCC) method for brain signals, a procedure for characterizing connectivity in a network by clustering nodes or groups of channels that display a high level of coordination as measured by “cluster-coherence.” While the most common approach to measure dependence between clusters is through pairs of single time series, our method proposes cluster coherence which measures dependence between pairs of whole clusters rather than between single elements. Thus it takes into account both the dependence between clusters and within channels in a cluster. The identified clusters contain time series that exhibit high cross-dependence in the spectral domain. Simulation studies demonstrate that the proposed HCC method is competitive with the other feature-based clustering methods. To study clustering in a network of multichannel electroencephalograms (EEG) during an epileptic seizure, we applied the HCC method and identified connectivity on alpha (8, 12) Hertz and beta (16, 30) Hertz bands at different phases of the recording: before an epileptic seizure, during the early and middle phases of the seizure episode. To increase the potential impact of HCC in neuroscience, we also developed the HCC-Vis, an R-Shiny app (RStudio), which can be downloaded from <https://carolinaeuan.shinyapps.io/hcc-vis/>.

REFERENCES

- BABADI, B. and BROWN, E. N. (2014). A review of multitaper spectral analysis. *IEEE Trans. Biomed. Eng.* **61** 1555–1564.
- BOWYER, S. M. (2016). Coherence a measure of the brain networks: Past and present. *Neuropsychiatric Electrophysiology* **2** 1.
- BRILLINGER, D. R. (1975). *Time Series: Data Analysis and Theory*. International Series in Decision Processes. Holt, Rinehart and Winston, New York. [MR0443257](#)
- BROCKWELL, P. J. and DAVIS, R. A. (2006). *Time Series: Theory and Methods*. Springer Series in Statistics. Springer, New York. Reprint of the second (1991) edition. [MR2839251](#)
- BUZSÁKI, G. and DRAGUHN, A. (2004). Neuronal oscillations in cortical networks. *Science* **304** 1926–1929.
- CAIADO, J., CRATO, N. and PEÑA, D. (2006). A periodogram-based metric for time series classification. *Comput. Statist. Data Anal.* **50** 2668–2684. [MR2227325](#)
- CAIADO, J., MAHARAJ, E. A. and D’URSO, P. (2016). Time-series clustering. In *Handbook of Cluster Analysis*. Chapman & Hall/CRC Handb. Mod. Stat. Methods 241–263. CRC Press, Boca Raton, FL. [MR3644715](#)

Key words and phrases. Cluster coherence, multivariate time series, electroencephalograms, spectral analysis, classification.

- CRIBBEN, I. and YU, Y. (2017). Estimating whole-brain dynamics by using spectral clustering. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 607–627. [MR3632344](#)
- D'URSO, P., DE GIOVANNI, L. and MASSARI, R. (2016). GARCH-based robust clustering of time series. *Fuzzy Sets and Systems* **305** 1–28. [MR3557847](#)
- D'URSO, P. and MAHARAJ, E. A. (2009). Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems* **160** 3565–3589. [MR2563306](#)
- D'URSO, P. and MAHARAJ, E. A. (2012). Wavelets-based clustering of multivariate time series. *Fuzzy Sets and Systems* **193** 33–61. [MR2880740](#)
- DECKER, S., FILLMORE, P. T. and COHERENCE, A. R. (2017). The measurement and application of brain connectivity. *NeuroRegulation* **4** 3–13.
- EUÁN, C., OMBAO, H. and ORTEGA, J. (2018). The hierarchical spectral merger algorithm: A new time series clustering procedure. *J. Classification* **35** 71–99. [MR3790113](#)
- EUÁN, C., SUN, Y. and OMBAO, H. (2019a). Supplement to “Coherence-based time series clustering for statistical inference and visualization of brain connectivity.” DOI:[10.1214/18-AOAS1225SUPPA](#).
- EUÁN, C., SUN, Y. and OMBAO, H. (2019b). Supplement to “Coherence-based time series clustering for statistical inference and visualization of brain connectivity.” DOI:[10.1214/18-AOAS1225SUPPB](#).
- FIECAS, M. and OMBAO, H. (2011). The generalized shrinkage estimator for the analysis of functional connectivity of brain signals. *Ann. Appl. Stat.* **5** 1102–1125. [MR2840188](#)
- FIECAS, M., OMBAO, H., LINKLETTER, C., THOMPSON, W. and SANES, J. (2010). Functional connectivity: Shrinkage estimation and randomization test. *NeuroImage* **49** 3901–3915.
- GAO, X., SHAHBABA, B., FORTIN, N. and OMBAO, H. (2016). Evolutionary state-space model and its application to time-frequency analysis of local field potentials. Available at [arXiv:1610.07271](#).
- KIMES, P. K., LIU, Y., HAYES, D. N. and MARRON, J. S. (2017). Statistical significance for hierarchical clustering. *Biometrics* **73** 811–821. [MR3713115](#)
- LAFUENTE-REGO, B., D'URSO, P. and VILAR, J. A. (2018). Robust fuzzy clustering based on quantile autocovariances. *Statist. Papers* 1–56.
- MAHARAJ, E. A. and D'URSO, P. (2010). A coherence-based approach for the pattern recognition of time series. *Phys. A* **389** 3516–3537. [MR2659314](#)
- MAHARAJ, E. A. and D'URSO, P. (2011). Fuzzy clustering of time series in the frequency domain. *Inform. Sci.* **181** 1187–1211.
- MAHARAJ, E. A., D'URSO, P. and GALAGEDERA, D. U. A. (2010). Wavelet-based fuzzy clustering of time series. *J. Classification* **27** 231–275. [MR2726320](#)
- NEDUNGADI, A. G., DING, M. and RANGARAJAN, G. (2011). Block coherence: A method for measuring the interdependence between two blocks of neurobiological time series. *Biol. Cybernet.* **104** 197–207. [MR2795598](#)
- OMBAO, H. and VAN BELLEGEM, S. (2008). Evolutionary coherence of nonstationary signals. *IEEE Trans. Signal Process.* **56** 2259–2266. [MR2516630](#)
- OMBAO, H. C., RAZ, J. A., STRAWDERMAN, R. L. and VON SACHS, R. (2001). A simple generalised crossvalidation method of span selection for periodogram smoothing. *Biometrika* **88** 1186–1192. [MR1872229](#)
- PAWITAN, Y. (1996). Automatic estimation of the cross-spectrum of a bivariate time series. *Biometrika* **83** 419–432. [MR1439793](#)
- RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.
- ROSEN, O. and STOFFER, D. S. (2007). Automatic estimation of multivariate spectra via smoothing splines. *Biometrika* **94** 335–345. [MR2331489](#)
- SHUMWAY, R. H. and STOFFER, D. S. (2011). *Time Series Analysis and Its Applications. With R Examples*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. [MR2721825](#)

- TAKAHASHI, D. Y., BACCALÁ, L. A. and SAMESHIMA, K. (2014). Canonical information flow decomposition among neural structure subsets. *Frontiers in Neuroinformatics* **8** 49.
- TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 411–423. [MR1841503](#)
- VIQUEIRA, M., GARCÍA ZAPIRAIN, B. and MENDEZ ZORRILLA, A. (2013). Ocular movement and cardiac rhythm control using eeg techniques. In *Medical Imaging in Clinical Practice* (O. F. Eröndü, ed.), Chapter 6. InTech, Rijeka.
- WALDEN, A. T. (2000). A unified view of multitaper multivariate spectral estimation. *Biometrika* **87** 767–788. [MR1813974](#)

SPARSE PRINCIPAL COMPONENT ANALYSIS WITH MISSING OBSERVATIONS¹

BY SEYOUNG PARK AND HONGYU ZHAO

Sungkyunkwan University and Yale University

Principal component analysis (PCA) is a commonly used statistical method in a wide range of applications. However, it does not work well when the number of features is larger than the sample size. We consider the estimation of the sparse principal subspace in the high dimensional setting with missing data motivated by the analysis of single-cell RNA sequence data. We propose a two step estimation procedure, and establish the rates of convergence for estimating the principal subspace. Simulated examples with various missing mechanisms show its competitive performance compared to existing sparse PCA methods. We apply the method to single-cell data and show that the proposed method can better distinguish cell types than other PCA methods.

REFERENCES

- AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.* **40** 2452–2482. [MR3097609](#)
- AMINI, A. A. and WAINWRIGHT, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. [MR2541450](#)
- BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. [MR2279680](#)
- BAILEY, S. (2012). Principal component analysis with noisy and/or missing data. *Publications of the Astronomical Society of the Pacific* **124** 1015–1023.
- BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. [MR3127849](#)
- CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. [MR3161458](#)
- CAI, T. T. and ZHANG, A. (2016). Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *J. Multivariate Anal.* **150** 55–74. [MR3534902](#)
- CANNOODT, R. (2016). Scorpis improves trajectory inference and identifies novel modules in dendritic cell development. Preprint.
- CHEN, H. (2002). Principal component analysis with missing data and outliers. Robust Image Understanding Laboratory. Tech. report.
- DELCHAMBRE, L. (2014). Weighted principal component analysis: A weighted covariance eigendecomposition approach. *Mon. Not. R. Astron. Soc.* **446** 3545–3555.
- DENG, Q., RAMSKÖLD, D., REINIUS, B. and SANDBERG, R. (2014). Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343** 193–196.
- DESHPANDE, Y. and MONTANARI, A. (2016). Sparse PCA via covariance thresholding. *J. Mach. Learn. Res.* **17** 1–41. [MR3555032](#)

Key words and phrases. PCA, missing data, high dimensional, single-cell data.

- DODGE, Y. (1985). *Analysis of Experiments with Missing Data. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, Chichester. [MR0811638](#)
- ECKART, C. and YOUNG, G. (1936). The approximation of one matrix by another of low rank. *Psychometrika* **1** 211.
- FORGY, E. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics* **21** 768–769.
- HASTIE, T., MAZUMDER, R., LEE, J. D. and ZADEH, R. (2015). Matrix completion and low-rank SVD via fast alternating least squares. *J. Mach. Learn. Res.* **16** 3367–3402. [MR3450542](#)
- HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24** 417–441.
- JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. [MR2751448](#)
- JOLLIFFE, I. T. (1986). *Principal Component Analysis. Springer Series in Statistics*. Springer, New York. [MR0841268](#)
- JOURNÉE, M., NESTEROV, Y., RICHTÁRIK, P. and SEPULCHRE, R. (2010). Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.* **11** 517–553. [MR2600619](#)
- KUNDU, A., DRINEAS, P. and MAGDON-ISMAIL, M. (2015). Approximating sparse pca from incomplete data. *Adv. Neural Inf. Process. Syst.* **28** 388–396.
- LANGE, K., HUNTER, D. R. and YANG, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graph. Statist.* **9** 1–59. [MR1819865](#)
- LIN, P., TROUP, M. and HO, J. W. K. (2017). Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome Biol.* **18**.
- LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664. [MR3015038](#)
- LOUNICI, K. (2013). Sparse principal component analysis with missing observations. In *High Dimensional Probability VI. Progress in Probability* **66** 327–356. Birkhäuser/Springer, Basel. [MR3443508](#)
- MA, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. [MR3099121](#)
- MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)* 281–297. Univ. California Press, Berkeley, CA. [MR0214227](#)
- MAZUMDER, R., HASTIE, T. and TIBSHIRANI, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11** 2287–2322. [MR2719857](#)
- NADLER, B. (2009). Discussion of “On consistency and sparsity for principal components analysis in high dimensions,” by I. M. Johnstone and A. Y. Lu” [MR2751448]. *J. Amer. Statist. Assoc.* **104** 694–697. [MR2751449](#)
- PARK, S. and ZHAO, H. (2018). Spectral clustering based on learning similarity matrix. *Bioinformatics* **34** 2069–2076.
- PARK, S. and ZHAO, H. (2019). Supplement to “Sparse Principal Component Analysis with Missing Observations.” DOI:[10.1214/18-AOAS1220SUPP](#).
- PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. [MR2399865](#)
- PAUL, D. and JOHNSTONE, I. M. (2007). Augmented sparse principal component analysis for high dimensional data. Available at [arXiv:1202.1242v1](#).
- PIERSON, E. and YAU, C. (2015). Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16** 1–10.
- POLLEN, A., NOWAKOWSKI, T., SHUGA, J., WANG, X., LEYRAT, A., LUI, J., LI, N., SZPANKOWSKI, L., FOWLER, B., CHEN, P., RAMALINGAM, N., SUN, G., THU, M., NORRIS, M., LEBOWSKY, R., TOPPANI, D., KEMP, D., WONG, M., CLERKSON, B., JONES, B., WU, S.,

- KNUTSSON, L., ALVARADO, B., WANG, J., WEAVER, L., MAY, A., JONES, R., UNGER, M., KRIEGSTEIN, A. and WEST, J. (2014). Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32** 1053–1058.
- QI, X., LUO, R. and ZHAO, H. (2013). Sparse principal component analysis by choice of norm. *J. Multivariate Anal.* **114** 127–160. [MR2993878](#)
- SCHLITZER, A., SIVAKAMASUNDARI, V., CHEN, J., SUMATOH, H. R. B., SCHREUDER, J., LUM, J., MALLERET, B., ZHANG, S., LARBI, A., ZOLEZZI, F. et al. (2015). Identification of cdc1- and cdc2-committed dc progenitors reveals early lineage priming at the common dc progenitor stage in the bone marrow. *Nature Immunology* **16** 718–728.
- SHAO, C. and HÖFER, T. (2017). Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics* **33** 235–242.
- SREBRO, N., RENNIE, J. and JAAKKOLA, T. (2005). Maximum margin matrix factorization. *Adv. Neural Inf. Process. Syst.* **17**.
- STREHL, A. and GHOSH, J. (2003). Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3** 583–617. Computational learning theory. [MR1991087](#)
- TING, D. T., WITTNER, B. S., LIGORIO, M., JORDAN, N. V., SHAH, A. M., MIYAMOTO, D. T., ACETO, N., BERSANI, F., BRANNIGAN, B. W. et al. (2014). Single-cell rna sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Reports* **8** 1905–1918.
- TSALMANTZA, P. and HOGG, D. W. (2012). A data-driven model for spectra: Finding double redshifts in the sloan digital sky survey. *Astrophys. J.* **753** 1–16.
- USOSKIN, D., FURLAN, A., ISLAM, S., ABDO, H., LÖNNERBERG, P., LOU, D., HJERLING-LEFFLER, J., HAEGGSTRÖM, J., KHARCHENKO, O., KHARCHENKO, P. V., LINNARSSON, S. and ERNFORS, P. (2014). Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nat. Neurosci.* **18** 145–153.
- VU, V. Q. and LEI, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947. [MR3161452](#)
- WAGNER, S. and WAGNER, D. (2007). Comparing clusterings: An overview. Universität Karlsruhe, Fakultät Für Informatik Karlsruhe.
- WANG, Z., LU, H. and LIU, H. (2014). Tighten after relax: Minimax-optimal sparse pca in polynomial time. *Adv. Neural Inf. Process. Syst.* 3383–3391.
- WANG, B., ZHU, J., PIERSON, E., RAMAZZOTTI, D. and BATZOGLOU, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14** 414–416.
- WINTER, D. R. and AMIT, I. (2015). Dcs are ready to commit. *Nature Immunology* **16** 683–685.
- XU, C. and SU, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31** 1974–1980.
- YAN, L., YANG, M., GUO, H., YANG, L., WU, J., LI, R., LIU, P., LIAN, Y., ZHENG, X. et al. (2013). Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structure and Molecular Biology* **20** 1131–1142.
- YUAN, X.-T. and ZHANG, T. (2013). Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.* **14** 899–925. [MR3063614](#)
- ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. [MR2252527](#)

ADAPTIVE GPCA: A METHOD FOR STRUCTURED DIMENSIONALITY REDUCTION WITH APPLICATIONS TO MICROBIOME DATA

BY JULIA FUKUYAMA¹

Indiana University

Exploratory analysis is an important first step for discovering latent structure and generating hypotheses in large biological data sets. However, when the number of variables is large compared to the number of samples, standard methods such as principal components analysis give results that are unstable and difficult to interpret.

Here, we present adaptive generalized principal components analysis (adaptive gPCA), a new method that solves these problems by incorporating information about the relationships among the variables. Adaptive gPCA gives a low-dimensional representation of the samples with axes that are interpretable in terms of groups of closely related variables. We show that adaptive gPCA does well at reconstructing true latent structure in simulated data and demonstrate its use on a study of the effect of antibiotics on the human gut microbiota.

REFERENCES

- ALLEN, G. I., GROSENICK, L. and TAYLOR, J. (2014). A generalized least-square matrix decomposition. *J. Amer. Statist. Assoc.* **109** 145–159. [MR3180553](#)
- BRENNER, D. J., STALEY, J. T. and KRIEG, N. R. (2005). Classification of procaryotic organisms and the concept of bacterial speciation. In *Bergey's Manual of Systematic Bacteriology* 27–32. Springer, Berlin.
- CALLAHAN, B. J., SANKARAN, K., FUKUYAMA, J. A., MCMURDIE, P. J. and HOLMES, S. P. (2016). Bioconductor workflow for microbiome data analysis: From raw reads to community analyses. *F1000Res* **5** 1492.
- CAUSSINUS, H. (1986). Models and uses of principal component analysis. *Multidimensional Data Analysis* **86** 149–170.
- CHANG, Q., LUAN, Y. and SUN, F. (2011). Variance adjusted weighted unifracs: A powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinform.* **12** 1.
- CHANG, W., CHENG, J., ALLAIRE, J., XIE, Y. and MCPHERSON, J. (2016). shiny: Web Application Framework for R. R package version 0.13.2.
- CHEN, J., BITTINGER, K., CHARLSON, E. S., HOFFMANN, C., LEWIS, J., WU, G. D., COLLAMAN, R. G., BUSHMAN, F. D. and LI, H. (2012). Associating microbiome composition with environmental covariates using generalized unifracs distances. *Bioinformatics* **28** 2106–2113.
- COHAN, F. M. (2002). What are bacterial species? *Annual Reviews in Microbiology* **56** 457–487.
- 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.

Key words and phrases. PCA, microbiome, phylogeny, empirical Bayes.

- DOOLITTLE, W. F. and PAPKE, R. T. (2006). Genomics and the bacterial species problem. *Genome Biol.* **7** 1.
- DRAY, S., PAVOINE, S. and AGUIRRE DE CÁRCER, D. (2015). Considering external information to improve the phylogenetic comparison of microbial communities: A new approach based on constrained double principal coordinates analysis (cdpcoa). *Molecular Ecology Resources* **15** 242–249.
- EDGAR, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics* **26** 2460–2461.
- ESCOUFIER, Y. (1973). Le traitement des variables vectorielles. *Biometrics* **29** 751–760. [MR0334416](#)
- FERNANDES, A. D., REID, J. N., MACKLAIM, J. M., MCMURROUGH, T. A., EDGELL, D. R. and GLOOR, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: Characterizing rna-seq, 16s rna gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2** 15.
- FILZMOSER, P., HRON, K. and REIMANN, C. (2009). Principal component analysis for compositional data with outliers. *Environmetrics* **20** 621–632. [MR2838477](#)
- FUKUYAMA, J. (2019). Supplement to “Adaptive gPCA: A method for structured dimensionality reduction with applications to microbiome data.” DOI:[10.1214/18-AOAS1227SUPP](#).
- HOLMES, S. (2008). Multivariate data analysis: The French way. In *Probability and Statistics: Essays in Honor of David A. Freedman. Inst. Math. Stat. (IMS) Collect.* **2** 219–233. IMS, Beachwood, OH. [MR2459953](#)
- JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. [MR2751448](#)
- KONDOR, R. I. and LAFFERTY, J. (2002). Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th International Conference on Machine Learning* 315–322.
- LI, C. and LI, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24** 1175–1182.
- 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.
- LOZUPONE, C. and KNIGHT, R. (2005). Unifrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71** 8228–8235.
- LOZUPONE, C. A., HAMADY, M., KELLEY, S. T. and KNIGHT, R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology* **73** 1576–1585.
- MATSEN, F. A. and EVANS, S. N. (2013). Edge principal components and squash clustering: Using the special structure of phylogenetic placement data for sample comparison. *PLoS ONE*.
- MCMURDIE, P. J. and HOLMES, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10** e1003531.
- PARADIS, E., CLAUDE, J. and STRIMMER, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20** 289–290.
- PAVOINE, S., DUFOUR, A.-B. and CHESSEL, D. (2004). From dissimilarities among species to dissimilarities among communities: A double principal coordinate analysis. *J. Theoret. Biol.* **228** 523–537. [MR2080909](#)
- PENROSE, R. (1955). A generalized inverse for matrices. *Proc. Camb. Philos. Soc.* **51** 406–413. [MR0069793](#)
- PURDOM, E. (2011). Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree. *Ann. Appl. Stat.* **5** 2326–2358. [MR2907117](#)
- QUAST, C., PRUESSE, E., YILMAZ, P., GERKEN, J., SCHWEER, T., YARZA, P., PEPLIES, J. and GLÖCKNER, F. O. (2013). The silva ribosomal rna gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41** D590–D596.

- R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- RANDOLPH, T. W., ZHAO, S., COPELAND, W., HULLAR, M. and SHOJAIE, A. (2018). Kernel-penalized regression for analysis of microbiome data. *Ann. Appl. Stat.* **12** 540–566. [MR3773404](#)
- RAPAPORT, F., ZINOVYEV, A., DUTREIX, M., BARILLOT, E. and VERT, J.-P. (2007). Classification of microarray data using gene networks. *BMC Bioinform.* **8** 35.
- RINALDO, A. (2009). Properties and refinements of the fused lasso. *Ann. Statist.* **37** 2922–2952. [MR2541451](#)
- SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102** 15545–15550.
- TIBSHIRANI, R. and WANG, P. (2008). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics* **9** 18–29.
- TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. [MR2136641](#)
- WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

LEARNING ALGORITHMS TO EVALUATE FORENSIC GLASS EVIDENCE

BY SOYOUNG PARK AND ALICIA CARRIQUIRY¹

Iowa State University

Glass fragments are often compared in the course of a forensic evaluation using their chemical composition determined with technologies such as LA-ICP-MS. At present forensic scientists advocate the use of two comparison criteria based on univariate intervals around all mean elemental concentrations for fragments originating from a known piece of broken glass. The main drawback of this approach is that it does not consider the dependencies between concentrations. Further, when the elemental concentrations are more variable within panes, it becomes harder to reject the null hypothesis of no difference between fragments. In the legal context higher variance would tend to incriminate the defendant because the intervals would tend to be wider. We demonstrate that a score-based approach to assess the probative value of evidence in glass comparisons outperforms the two standard interval methods and other methods proposed in the literature, at least in terms of minimizing classification error in the glass fragment sources we analyzed. We use machine learning algorithms to construct a *similarity score* between pairs of glass fragments. The learning algorithms exploit the dependencies among elemental concentrations and result in an empirical class probability; so, we can report the *degree of similarity* between two fragments. Our group is in the process of assembling the first glass composition database with enough information within and between glass samples to permit computing well-conditioned estimates of high-dimensional covariance matrices. These data will be available to anyone who wishes to carry out research in this area.

REFERENCES

- AESCHLIMAN, D. B., BAJIC, S. J., BALDWIN, D. P. and HOUK, R. (2003). Spatially-resolved analysis of solids by laser ablation-inductively coupled plasma-mass spectrometry: Trace elemental quantification without matrix-matched solid standards. *J. Anal. At. Spectrom.* **18** 872–877.
- AITKEN, C. G. G. and LUCY, D. (2004). Evaluation of trace evidence in the form of multivariate data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **53** 109–122. [MR2037883](#)
- ASTM-E2330-12 (2012). Standard test method for determination of concentrations of elements in glass samples using inductively coupled plasma mass spectrometry (ICP-MS) for forensic comparisons. Technical report, ASTM International. Available at <https://doi.org/10.1520/E2330-12>.
- ASTM-E2927-16 (2016). Standard test method for determination of trace elements in soda-lime glass samples using laser ablation inductively coupled plasma mass spectrometry for forensic comparisons. Technical report, ASTM International. Available at <https://doi.org/10.1520/E2927-16>.

Key words and phrases. Multivariate measurements, random forest, out-of-bag errors, score likelihood ratio, forensic glass comparisons.

- BIRNBAUM, A. (1954). Combining independent tests of significance. *J. Amer. Statist. Assoc.* **49** 559–574. [MR0065101](#)
- BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.
- CAMPBELL, G. P. and CURRAN, J. M. (2009). The interpretation of elemental composition measurements from forensic glass evidence III. *Sci. Justice* **49** 2–7.
- CHAWLA, N. V., BOWYER, K. W., HALL, L. O. and KEGELMEYER, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *J. Artificial Intelligence Res.* **16** 321–357.
- CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. [MR2758172](#)
- CURRAN, J. M. (2003). The statistical interpretation of forensic glass evidence. *Int. Stat. Rev.* **71** 497–520.
- CURRAN, J. M., CHAMPOD, T. N. H. and BUCKLETON, J. S. (2000). *Forensic Interpretation of Glass Evidence*. CRC Press, Boca Raton, FL.
- CURRAN, J., TRIGGS, C., ALMIRALL, J., BUCKLETON, J. and WALSH, K. (1997a). The interpretation of elemental composition measurements from forensic glass evidence: II. *Sci. Justice* **37** 245–249.
- CURRAN, J., TRIGGS, C., ALMIRALL, J., BUCKLETON, J. and WALSH, K. (1997b). The interpretation of elemental composition measurements from forensic glass evidence: I. *Sci. Justice* **37** 241–244.
- DAVIS, L. J., SAUNDERS, C. P., HEPLER, A. and BUSCAGLIA, J. (2012). Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios. *Forensic Sci. Int.* **216** 146–157.
- DELONG, E. R., DELONG, D. M. and CLARKE-PEARSON, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44** 837–845.
- DORN, H., RUDDALL, D. E., HEYDON, A. and BURTON, B. D. (2015). Discrimination of float glass by LA-ICP-MS: Assessment of exclusion criteria using casework samples. *Can. Soc. Forensic Sci. J.* **48** 85–96.
- GARVIN, E. J. and KOONS, R. D. (2011). Evaluation of match criteria used for the comparison of refractive index of glass fragments. *J. Forensic Sci.* **56** 491–500.
- HARE, E., HOFMANN, H. and CARRIQUIRY, A. (2017). Automatic matching of bullet land impressions. *Ann. Appl. Stat.* **11** 2332–2356. [MR3743299](#)
- HEPLER, A. B., SAUNDERS, C. P., DAVIS, L. J. and BUSCAGLIA, J. (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Sci. Int.* **219** 129–140.
- HICKMAN, D. (1987). Glass types identified by chemical analysis. *Forensic Sci. Int.* **33** 23–46.
- HOUK, R. S. (1990). Elemental analysis by atomic emission and mass spectrometry with inductively coupled plasmas. In *Handbook on the Physics and Chemistry of Rare Earths* **13** 385–421.
- KAM, H. T. (1995). Random decision forest. In *Proc. of the 3rd Int'l Conf. on Document Analysis and Recognition, Montreal, Canada, August 14–18*.
- KAPELNER, A. and BLEICH, J. (2013). bartMachine: Machine learning with Bayesian additive regression trees. Preprint. Available at [arXiv:1312.2171](#).
- KOONS, R. D. and BUSCAGLIA, J. (2002). Interpretation of glass composition measurements: The effects of match criteria on discrimination capability. *J. Forensic Sci.* **47** 505–512.
- KOONS, R. D., FIEDLER, C. and RAWALT, R. (1988). Classification and discrimination of sheet and container glasses by inductively coupled plasma-atomic emission spectrometry and pattern recognition. *J. Forensic Sci.* **33** 49–67.
- KOONS, R. D., PETERS, C. A. and REBBERT, P. S. (1991). Comparison of refractive index, energy dispersive X-ray fluorescence and inductively coupled plasma atomic emission spectrometry for forensic characterization of sheet glass fragments. *J. Anal. At. Spectrom.* **6** 451–456.
- LEDOIT, O. and WOLF, M. (2003). Honey, I shrunk the sample covariance matrix. *J. Portfolio Management* **30**.

- LINDLEY, D. V. (1977). A problem in forensic science. *Biometrika* **64** 207–213. [MR0518935](#)
- LUNARDON, N., MENARDI, G. and TORELLI, N. (2014). ROSE: A package for binary imbalanced learning. *R J.* **6** 79–89.
- LUND, S. P. and IYER, H. K. (2017). Likelihood ratio as weight of forensic evidence: A closer look. Preprint. Available at [arXiv:1704.08275](#).
- MORRISON, G. S. and ENZINGER, E. (2016). What should a forensic practitioner's likelihood ratio be? *Sci. Justice* **56** 374–379.
- OMMEN, D. M. and SAUNDERS, C. P. (2018). Building a unified statistical framework for the forensic identification of source problems. *Law Probab. Risk* **17** 179–197.
- PARKER, J. B. (1966). A statistical treatment of identification problems. *Sci. Justice* **6** 33–39.
- PARKER, J. B. (1967). The mathematical evaluation of numerical evidence. *Sci. Justice* **7** 134–144.
- PARKER, J. B. and HOLFORD, A. (1968). Optimum test statistics with particular reference to a forensic science problem. *Appl. Stat.* **17** 237–251.
- SCHÄFER, J. and STRIMMER, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 32, 28. [MR2183942](#)
- SCHEER, S. (2006). The evidential value of elemental composition in forensic glass examination: The use of multivariate likelihood ratio methods. Available at <https://www.forensic.to/webhome/statistiek/AfstudeerscriptieSonja.pdf>.
- SONG, J. (2015). Proposed congruent matching cells (CMC) method for ballistic identification and error rate estimation. *AFTE J.* **3** 177–185.
- SWOFFORD, H., KOERTNER, A., ZEMP, F., AUSDEMORE, M., LIU, A. and SALYARDS, M. (2018). A method for the statistical interpretation of friction ridge skin impression evidence: Method development and validation. *Forensic Sci. Int.* **287** 113–126.
- TANGRAM-TECHNOLOGY (2004). Float glass production. Available at <http://www.tangram.co.uk/TI-Glazing-Float%20Glass.html>.
- TREJOS, T., FLORES, A. and ALMIRALL, J. R. (2010). Micro-spectrochemical analysis of document paper and gel inks by laser ablation inductively coupled plasma mass spectrometry and laser induced breakdown spectroscopy. *Spectrochim. Acta, Part B: Atom. Spectrosc.* **65** 884–895.
- TREJOS, T., KOONS, R., BECKER, S., BERMAN, T., BUSCAGLIA, J., DUECKING, M., ECKERT-LUMSDON, T., ERNST, T., HANLON, C., HEYDON, A. et al. (2013a). Cross-validation and evaluation of the performance of methods for the elemental analysis of forensic glass by μ -XRF, ICP-MS, and LA-ICP-MS. *Anal. Bioanal. Chem.* **405** 5393–5409.
- TREJOS, T., KOONS, R., WEIS, P., BECKER, S., BERMAN, T., DALPE, C., DUECKING, M., BUSCAGLIA, J., ECKERT-LUMSDON, T., ERNST, T. et al. (2013b). Forensic analysis of glass by μ -XRF, SN-ICP-MS, LA-ICP-MS and LA-ICP-OES: Evaluation of the performance of different criteria for comparing elemental composition. *J. Anal. At. Spectrom.* **28** 1270–1282.
- WEIS, P., DÜCKING, M., WATZKE, P., MENGES, S. and BECKER, S. (2011). Establishing a match criterion in forensic comparison analysis of float glass using laser ablation inductively coupled plasma mass spectrometry. *J. Anal. At. Spectrom.* **26** 1273–1284.
- ZADORA, G. (2009a). Classification of glass fragments based on elemental composition and refractive index. *J. Forensic Sci.* **54** 49–59.
- ZADORA, G. (2009b). Evaluation of evidence value of glass fragments by likelihood ratio and Bayesian network approaches. *Anal. Chim. Acta* **642** 279–290.

THREE-WAY CLUSTERING OF MULTI-TISSUE MULTI-INDIVIDUAL GENE EXPRESSION DATA USING SEMI-NONNEGATIVE TENSOR DECOMPOSITION¹

BY MIAOYAN WANG*, JONATHAN FISCHER[†] AND YUN S. SONG[†]

University of Wisconsin—Madison and University of California, Berkeley[†]*

The advent of high-throughput sequencing technologies has led to an increasing availability of large multi-tissue data sets which contain gene expression measurements across different tissues and individuals. In this setting, variation in expression levels arises due to contributions specific to genes, tissues, individuals, and interactions thereof. Classical clustering methods are ill-suited to explore these three-way interactions and struggle to fully extract the insights into transcriptome complexity contained in the data. We propose a new statistical method, called *MultiCluster*, based on semi-nonnegative tensor decomposition which permits the investigation of transcriptome variation across individuals and tissues simultaneously. We further develop a tensor projection procedure which detects covariate-related genes with high power, demonstrating the advantage of tensor-based methods in incorporating information across similar tissues. Through simulation and application to the GTEx RNA-seq data from 53 human tissues, we show that *MultiCluster* identifies three-way interactions with high accuracy and robustness.

REFERENCES

- ALLEN, G. (2012). Sparse higher-order principal components analysis . In *Proc. Fifteenth International Conference on Artificial Intelligence and Statistics*, PMLR 27–36. Available at <http://proceedings.mlr.press/v22/allen12.html>.
- ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15** 2773–2832. [MR3270750](#)
- BAHCALL, O. G. (2015). Human genetics: GTEx pilot quantifies eQTL variation across tissues and individuals. *Nat. Rev. Genet.* **16** 375.
- BATTLE, A., BROWN, C. D., ENGELHARDT, B. E., MONTGOMERY, S. B., CONSORTIUM, G. et al. (2017). Genetic effects on gene expression across human tissues. *Nature* **550** 204–213.
- 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](#)
- CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. [MR2565240](#)
- CARRASQUILLO, M. M., ZOU, F., PANKRATZ, V. S., WILCOX, S. L., MA, L., WALKER, L. P., YOUNKIN, S. G., YOUNKIN, C. S., YOUNKIN, L. H., BISCEGLIO, G. D. et al. (2009). Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer’s disease. *Nat. Genet.* **41** 192–198.
- GTEx CONSORTIUM (2015). The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348** 648–660.

Key words and phrases. Tensor decomposition, tensor projection, clustering, gene expression.

- DE LATHAUWER, L. (2006). A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM J. Matrix Anal. Appl.* **28** 642–666. [MR2262974](#)
- DE SILVA, V. and LIM, L.-H. (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM J. Matrix Anal. Appl.* **30** 1084–1127. [MR2447444](#)
- DEY, K. K., HSIAO, C. J. and STEPHENS, M. (2017). Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genet.* **13** e1006599.
- FISHILEVICH, S., ZIMMERMAN, S., KOHN, A., STEIN, T. I., OLENDER, T., KOLKER, E., SAFRAN, M. and LANCET, D. (2016). Genic insights from integrated human proteomics in GeneCards. *Database (Oxford)* **2016**.
- GAO, C., MCDOWELL, I. C., ZHAO, S., BROWN, C. D. and ENGELHARDT, B. E. (2016). Context specific and differential gene co-expression networks via Bayesian biclustering. *PLoS Comput. Biol.* **12** e1004791.
- HAWRYLYCZ, M. J., LEIN, S., GUILLOZET-BONGAARTS, A. L., SHEN, E. H., NG, L., MILLER, J. A., VAN DE LAGEMAAT, L. N., SMITH, K. A., EBBERT, A., RILEY, Z. L. et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489** 391.
- HILLAR, C. J. and LIM, L.-H. (2013). Most tensor problems are NP-hard. *J. ACM* **60** Art. 45, 39. [MR3144915](#)
- HITCHCOCK, F. L. (1927). The expression of a tensor or a polyadic as a sum of products. *Stud. Appl. Math.* **6** 164–189.
- HORE, V., VIÑUELA, A., BUIL, A., KNIGHT, J., MCCARTHY, M. I., SMALL, K. and MARCHINI, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nat. Genet.* **48** 1094–1100.
- KELLEY, G. A. and KELLEY, K. S. (2012). Statistical models for meta-analysis: A brief tutorial. *World J. Methodol.* **2** 27–32.
- KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. [MR2535056](#)
- KRUSKAL, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.* **18** 95–138. [MR0444690](#)
- LAM, A. D., DECK, G., GOLDMAN, A., ESKANDAR, E. N., NOEBELS, J. and COLE, A. J. (2017). Silent hippocampal seizures and spikes identified by foramen ovale electrodes in Alzheimer's disease. *Nat. Med.* **23** 678–680.
- LAZZERONI, L. and OWEN, A. (2002). Plaid models for gene expression data. *Statist. Sinica* **12** 61–86. [MR1894189](#)
- LEE, S. and HUANG, J. Z. (2014). A biclustering algorithm for binary matrices based on penalized Bernoulli likelihood. *Stat. Comput.* **24** 429–441. [MR3192266](#)
- LEE, S., HUANG, J. Z. and HU, J. (2010). Sparse logistic principal components analysis for binary data. *Ann. Appl. Stat.* **4** 1579–1601. [MR2758342](#)
- LIM, L.-H. (2005). Singular values and eigenvalues of tensors: A variational approach. In 2005 1st *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing* 129–132. IEEE, New York.
- LIU, Y., HAYES, D. N., NOBEL, A. and MARRON, J. S. (2008). Statistical significance of clustering for high-dimension, low-sample size data. *J. Amer. Statist. Assoc.* **103** 1281–1293. [MR2528840](#)
- LOCK, E. F., HOADLEY, K. A., MARRON, J. S. and NOBEL, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7** 523–542. [MR3086429](#)
- LONSDALE, J., THOMAS, J., SALVATORE, M., PHILLIPS, R., LO, E., SHAD, S., HASZ, R., WALTERS, G., GARCIA, F., YOUNG, N. et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45** 580–585.

- MELÉ, M., FERREIRA, P. G., REVERTER, F., DELUCA, D. S., MONLONG, J., SAMMETH, M., YOUNG, T. R., GOLDMANN, J. M., PERVOUCHINE, D. D., SULLIVAN, T. J. et al. (2015). The human transcriptome across tissues and individuals. *Science* **348** 660–665.
- MORI, F., TANJI, K., MIKI, Y., TOYOSHIMA, Y., YOSHIDA, M., KAKITA, A., TAKAHASHI, H., UTSUMI, J., SASAKI, H. and WAKABAYASHI, K. (2016). G protein-coupled receptor 26 immunoreactivity in intranuclear inclusions associated with polyglutamine and intranuclear inclusion body diseases. *Neuropathology* **36** 50–55.
- MU, C., HSU, D. and GOLDFARB, D. (2015). Successive rank-one approximations for nearly orthogonally decomposable symmetric tensors. *SIAM J. Matrix Anal. Appl.* **36** 1638–1659. [MR3432147](#)
- OMBERG, L., GOLUB, G. H. and ALTER, O. (2007). A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proc. Natl. Acad. Sci. USA* **104** 18371–18376.
- PIERSON, E., KOLLER, D., BATTLE, A., MOSTAFAVI, S., CONSORTIUM, G. et al. (2015). Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Comput. Biol.* **11** e1004220.
- PRIDDLE, T. H. and CROW, T. J. (2013). The protocadherin 11X/Y (PCDH11X/Y) gene pair as determinant of cerebral asymmetry in modern Homo sapiens. *Ann. N.Y. Acad. Sci.* **1288** 36–47.
- TIBSHIRANI, R., HASTIE, T., EISEN, M., ROSS, D., BOTSTEIN, D., BROWN, P. et al. (1999). Clustering methods for the analysis of DNA microarray data. Technical report, Dept. Statistics, Stanford Univ., Stanford, CA.
- TUCKER, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika* **31** 279–311. [MR0205395](#)
- VAN DER MAATEN, L. and HINTON, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** 2579–2605.
- VEERAPPA, A. M., SALDANHA, M., PADAKANNAYA, P. and RAMACHANDRA, N. B. (2013). Genome-wide copy number scan identifies disruption of PCDH11X in developmental dyslexia. *Am. J. Med. Genet., Part B Neuropsychiatr. Genet.* **162** 889–897.
- WANG, M., FISCHER, J. and SONG, Y. S. (2019). Supplement to “Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition.” DOI:[10.1214/18-AOAS1228SUPP](#).
- WANG, M. and SONG, Y. S. (2017). Tensor decompositions via two-mode higher-order SVD (HOSVD). In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research* **54** 614–622.
- WANG, M., DAO DUC, K., FISCHER, J. and SONG, Y. S. (2017). Operator norm inequalities between tensor unfoldings on the partition lattice. *Linear Algebra Appl.* **520** 44–66. [MR3611456](#)
- YANG, J., HUANG, T., PETRALIA, F., LONG, Q., ZHANG, B., ARGMANN, C., ZHAO, Y., MOBBS, C. V., SCHADT, E. E., ZHU, J. et al. (2015). Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. *Sci. Rep.* **5** 15145.

NONPARAMETRIC TESTING FOR DIFFERENCES IN ELECTRICITY PRICES: THE CASE OF THE FUKUSHIMA NUCLEAR ACCIDENT

BY DOMINIK LIEBL

University of Bonn

This work is motivated by the problem of testing for differences in the mean electricity prices before and after Germany's abrupt nuclear phaseout after the nuclear disaster in Fukushima Daiichi, Japan, in mid-March 2011. Taking into account the nature of the data and the auction design of the electricity market, we approach this problem using a Local Linear Kernel (LLK) estimator for the nonparametric mean function of sparse covariate-adjusted functional data. We build upon recent theoretical work on the LLK estimator and propose a two-sample test statistics using a finite sample correction to avoid size distortions. Our nonparametric test results on the price differences point to a Simpson's paradox explaining an unexpected result recently reported in the literature.

REFERENCES

- BALL, R. and BROWN, P. (1968). An empirical evaluation of accounting income numbers. *J. Acc. Res.* 159–178.
- BENKO, M., HÄRDLE, W. and KNEIP, A. (2009). Common functional principal components. *Ann. Statist.* **37** 1–34. [MR2488343](#)
- BENTH, F. E., KHOLODNYI, V. A. and LAURENCE, P., eds. (2014). *Quantitative Energy Finance: Modeling, Pricing, and Hedging in Energy and Commodity Markets*. Springer, New York. [MR3185034](#)
- BETZER, A., DOUMET, M. and RINNE, U. (2013). How policy changes affect shareholder wealth: The case of the Fukushima Daiichi nuclear disaster. *Appl. Econ. Lett.* **20** 799–803.
- BRUNINX, K., MADZHAROV, D., DELARUE, E. and D'HAESELEER, W. (2013). Impact of the German nuclear phase-out on Europe's electricity generation—a comprehensive study. *Energy Policy* **60** 251–261.
- BUBLITZ, A., KELES, D. and FICHTNER, W. (2017). An analysis of the decline of electricity spot prices in Europe: Who is to blame? *Energy Policy* **107** 323–336.
- BURGER, M., GRAEBER, B. and SCHINDLMAYR, G. (2008). *Managing Energy Risk: An Integrated View on Power and Other Energy Markets*, 1st ed. Wiley, New York.
- BURGER, M., KLAR, B., MÜLLER, A. and SCHINDLMAYR, G. (2004). A spot market model for pricing derivatives in electricity markets. *Quant. Finance* **4** 109–122.
- CAO, G., YANG, L. and TODEM, D. (2012). Simultaneous inference for the mean function based on dense functional data. *J. Nonparametr. Stat.* **24** 359–377. [MR2921141](#)
- CARDOT, H. (2007). Conditional functional principal components analysis. *Scand. J. Stat.* **34** 317–335. [MR2346642](#)

Key words and phrases. Electricity spot prices, functional data analysis, local linear kernel estimation, nuclear power phaseout, sparse functional data, time series analysis.

- CHARNIGO, R. and SRINIVASAN, C. (2015). A multivariate generalized C_p and surface estimation. *Biostatistics* **16** 311–325. [MR3365430](#)
- CLUDIUS, J., HERMANN, H., MATTHES, F. C. and GRAICHEN, V. (2014). The merit order effect of wind and photovoltaic electricity generation in Germany 2008–2016: Estimation and distributional implications. *Energy Econ.* **44** 302–313.
- COX, D. D. and LEE, J. S. (2008). Pointwise testing with functional data using the Westfall–Young randomization method. *Biometrika* **95** 621–634. [MR2443179](#)
- FAMA, E. F., FISHER, L., JENSEN, M. C. and ROLL, R. (1969). The adjustment of stock prices to new information. *Internat. Econom. Rev.* **10** 1–21.
- FAN, J. and YAO, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. *Springer Series in Statistics*. Springer, New York. [MR1964455](#)
- FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. *Springer Series in Statistics*. Springer, New York. [MR2229687](#)
- FERSTL, R., UTZ, S. and WIMMER, M. (2012). The effect of the Japan 2011 disaster on nuclear and alternative energy stocks worldwide: An event study. *Bus. Res.* **5** 25–41.
- FOGARTY, C. B. and SMALL, D. S. (2014). Equivalence testing for functional data with an application to comparing pulmonary function devices. *Ann. Appl. Stat.* **8** 2002–2026. [MR3292487](#)
- GROMENKO, O. and KOKOSZKA, P. (2012). Testing the equality of mean functions of ionospheric critical frequency curves. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 715–731. [MR2993506](#)
- GROMENKO, O., KOKOSZKA, P. and SOJKA, J. (2017). Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves. *Ann. Appl. Stat.* **11** 898–918. [MR3693551](#)
- GROSSI, L., HEIM, S. and WATERSON, M. (2017). The impact of the German response to the Fukushima earthquake. *Energy Econ.* **66** 450–465.
- HALL, P., MÜLLER, H.-G. and WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. [MR2278365](#)
- HALL, P. and VAN KEILEGOM, I. (2007). Two-sample tests in functional data analysis starting from discrete data. *Statist. Sinica* **17** 1511–1531. [MR2413533](#)
- HIRTH, L. (2013). The market value of variable renewables: The effect of solar wind power variability on their relative price. *Energy Econ.* **38** 218–236.
- HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. *Springer Series in Statistics*. Springer, New York. [MR2920735](#)
- HORVÁTH, L., KOKOSZKA, P. and REEDER, R. (2013). Estimation of the mean of functional time series and a two-sample problem. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 103–122. [MR3008273](#)
- HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. [MR3379106](#)
- JIANG, C.-R. and WANG, J.-L. (2010). Covariate adjusted functional principal components analysis for longitudinal data. *Ann. Statist.* **38** 1194–1226. [MR2604710](#)
- LI, M., STAIU, A.-M. and BONDELL, H. D. (2015). Incorporating covariates in skewed functional data models. *Biostatistics* **16** 413–426. [MR3365436](#)
- LIEBL, D. (2013). Modeling and forecasting electricity spot prices: A functional data perspective. *Ann. Appl. Stat.* **7** 1562–1592. [MR3127959](#)
- LIEBL, D. (2019a). Inference for sparse and dense functional data with covariate adjustments. *J. Multivariate Anal.* **170** 315–335.
- LIEBL, D. (2019b). Supplement to “Nonparametric testing for differences in electricity prices: The case of the Fukushima nuclear accident.” DOI:[10.1214/18-AOAS1230SUPPA](#), DOI:[10.1214/18-AOAS1230SUPPB](#).
- LIESEN, M. G. (2007). The real-time price elasticity of electricity. *Energy Econ.* **29** 249–258.
- LISI, F. and SHAH, I. (2018). Forecasting next-day electricity demand and prices based on functional models. Working paper.

- MACIEJOWSKA, K. and WERON, R. (2016). Short-and mid-term forecasting of baseload electricity prices in the UK: The impact of intra-day price relationships and market fundamentals. *IEEE Trans. Power Syst.* **31** 994–1005.
- MACKINLAY, A. C. (1997). Event studies in economics and finance. *J. Econ. Lit.* **35** 13–39.
- MARTINS-FILHO, C. and YAO, F. (2007). Nonparametric frontier estimation via local linear regression. *J. Econometrics* **141** 283–319. [MR2411745](#)
- MCDERMOTT, G. R. and NILSEN, Ø. A. (2014). Electricity prices, river temperatures, and cooling water scarcity. *Land Econ.* **90** 131–148.
- MCWILLIAMS, A. and SIEGEL, D. (1997). Event studies in management research: Theoretical and empirical issues. *Acad. Manage. J.* **40** 626–657.
- PARASCHIV, F., ERNI, D. and PIETSCH, R. (2014). The impact of renewable energies on EEX day-ahead electricity prices. *Energy Policy* **73** 196–210.
- RABI, A., HADZIMA-NYARKO, M. and ŠPERAC, M. (2015). Modelling river temperature from air temperature: Case of the river Drava (Croatia). *Hydrol. Sci. J.* **60** 1490–1507.
- RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR2168993](#)
- ROHATGI, A. (2018). WebPlotDigitizer, version: 4.1. Austin, TX.
- SENSFUSS, F., RAGWITZ, M. and GENOESE, M. (2008). The merit-order effect: A detailed analysis of the price effect of renewable electricity generation on spot market prices in Germany. *Energy Policy* **36** 3086–3094.
- SERBAN, N. (2011). A space-time varying coefficient model: The equity of service accessibility. *Ann. Appl. Stat.* **5** 2024–2051. [MR2884930](#)
- SHAH, I. and LISI, F. (2018). Forecasting of electricity price through a functional prediction of supply and demand curves. Working paper.
- SHEATHER, S. J. and JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53** 683–690. [MR1125725](#)
- THOENES, S. (2014). Understanding the determinants of electricity prices and the impact of the German nuclear moratorium in 2011. *Energy J.* **35** 61–78.
- VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*, 4th ed. Springer, New York.
- VSEVOLOZHSKAYA, O., GREENWOOD, M. and HOLODOV, D. (2014). Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis. *Ann. Appl. Stat.* **8** 905–925. [MR3262539](#)
- WAGNER, C. H. (1982). Simpson's paradox in real life. *Amer. Statist.* **36** 46–48.
- YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. [MR2160561](#)
- ZHANG, J.-T. and CHEN, J. (2007). Statistical inferences for functional data. *Ann. Statist.* **35** 1052–1079. [MR2341698](#)
- ZHANG, X. and WANG, J.-L. (2016). From sparse to dense functional data and beyond. *Ann. Statist.* **44** 2281–2321. [MR3546451](#)
- ZHANG, W. and WEI, Y. (2015). Regression based principal component analysis for sparse functional data with applications to screening growth paths. *Ann. Appl. Stat.* **9** 597–620. [MR3371327](#)

NONPARAMETRIC INFERENCE FOR IMMUNE RESPONSE THRESHOLDS OF RISK IN VACCINE STUDIES¹

BY KEVIN M. DONOVAN*, MICHAEL G. HUDGENS* AND
 PETER B. GILBERT[†]

University of North Carolina at Chapel Hill and Fred Hutchinson Cancer
 Research Center[†]*

An important objective in vaccine studies entails identifying an immune response which is predictive of disease risk. Nonparametric methods are developed for inference on immune response thresholds that are associated with specified levels of disease risk, including where the risk level is zero. This threshold is defined as the minimum immune response value above which disease risk is less than or equal to the desired level. The proposed nonparametric methods are compared to previously developed parametric methods in simulation studies. The methods are extended for use in studies that only measure the immune response in a subset of participants, such as case-cohort or case-control studies, and with right censored time to disease outcomes. Finally, these methods are used to estimate neutralizing antibody thresholds for virologically confirmed dengue risk using data from two recent dengue vaccine trials.

REFERENCES

- ANDREWS, N., BORROW, R. and MILLER, E. (2003). Validation of serological correlate of protection for meningococcal c conjugate vaccine by using efficacy estimates from postlicensure surveillance in England. *Clin. Diagn. Lab. Immunol.* **10** 780–786.
- BLACK, S., NICOLAY, U., VESIKARI, T., KNUF, M., DEL GIUDICE, G., DELLA CIOPPA, G., TSAI, T., CLEMENS, R. and RAPPUOLI, R. (2011). Hemagglutination inhibition antibody titers as a correlate of protection for inactivated influenza vaccines in children. *Pediatr. Infect. Dis. J.* **30** 1081–1085.
- BORROW, R., BALMER, P. and MILLER, E. (2005). Meningococcal surrogates of protection—serum bactericidal antibody activity. *Vaccine* **23** 2222–2227.
- BRESLOW, N., LUMLEY, T., BALLANTYNE, C., CHAMBLESS, L. and KULICH, M. (2009). Using the whole cohort in the analysis of case-cohort data. *Am. J. Epidemiol.* **169** 1398–1405.
- CAPEDING, M. R., TRAN, N. H., HADINEGORO, S. R. S., ISMAIL, H. I. H. M., CHOTPITAYA-SUNONDH, T. and CHUA, M. N. (2014). Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy children in Asia: A phase 3, randomised, observer-masked, placebo-controlled trial. *Lancet* **384** 1358–1365.
- CENTERS FOR DISEASE CONTROL AND PREVENTION (2014). What would happen if we stopped vaccinations? Available at <https://www.cdc.gov/vaccines/vac-gen/whatifstop.htm>. Accessed June, 2018.
- CHAN, I. S. F., LI, S., MATTHEWS, H., CHAN, C., VESSEY, R., SADOFF, J. and HEYSE, J. (2002). Use of statistical models for evaluating antibody response as a correlate of protection against varicella. *Stat. Med.* **21** 3411–3430.

Key words and phrases. Case-cohort sampling, nonparametric, risk threshold, vaccine studies.

- CHEN, X., BAILLEUX, F., DESAI, K., QIN, L. and DUNNING, A. J. (2013). A threshold method for immunological correlates of protection. *BMC Med. Res. Methodol.* **13** 29.
- COLE, S. R., CHU, H. and GREENLAND, S. (2006). Multiple-imputation for measurement-error correction. *Int. J. Epidemiol.* **35** 1074–1081.
- COOKE, P. (1979). Statistical inference for bounds of random variables. *Biometrika* **66** 367–374. [MR0548207](#)
- CORNFIELD, J. (1951). A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *J. Natl. Cancer Inst.* **11** 1269–1275.
- DONOVAN, K. M., HUDGENS, M. G. and GILBERT, P. B. (2019). Supplement to “Nonparametric inference for immune response thresholds of risk in vaccine studies.” DOI:[10.1214/18-AOAS1237SUPP](#).
- DUNNING, A. J. (2006). A model for immunological correlates of protection. *Stat. Med.* **25** 1485–1497. [MR2227290](#)
- DYE, C. (2014). After 2015: Infectious diseases in a new era of health and development. *Phil. Trans. R. Soc. B* **369** 20130426.
- EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. [MR1270903](#)
- HADINEGORO, S. R., ARREDONDO-GARCÍA, J. L., CAPEDEING, M. R., DESEDA, C., CHOTPI-TAYASUNONDH, T. and DIETZE, R. (2015). Efficacy and long-term safety of a dengue vaccine in regions of endemic disease. *N. Engl. J. Med.* **373** 1195–1206.
- HAYNES, B. F., GILBERT, P. B., MCEL RATH, M. J., ZOLLA-PAZNER, S., TOMARAS, G. D., ALAM, S. M., EVANS, D. T., MONTEFIORI, D. C., KARNASUTA, C. et al. (2012). Immune-correlates analysis of an HIV-1 vaccine efficacy trial. *N. Engl. J. Med.* **366** 1275–1286.
- HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. [MR0053460](#)
- JÓDAR, L., BUTLER, J., CARLONE, G., DAGAN, R., GOLDBLATT, D. and KÄYHTY, H. (2003). Serological criteria for evaluation and licensure of new pneumococcal conjugate vaccine formulations for use in infants. *Vaccine* **21** 3265–3272.
- JOKINEN, J. T. and ÅHMAN, H. (2004). Concentration of antipneumococcal antibodies as a serological correlate of protection: An application to acute otitis media. *J. Infect. Dis.* **190** 545–550.
- MOODIE, Z., JURASKA, M., HUANG, Y., ZHUANG, Y., FONG, Y., CARPP, L. N., SELF, S. G., CHAMBONNEAU, L., SMALL, R. et al. (2018). Neutralizing antibody correlates analysis of tetravalent dengue vaccine efficacy trials in Asia and Latin America. *J. Infect. Dis.* **217** 742–753.
- PEPE, M. S. and FLEMING, T. R. (1989). Weighted Kaplan–Meier statistics: A class of distance tests for censored survival data. *Biometrics* **45** 497–507. [MR1010515](#)
- PRENTICE, R. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73** 1–11.
- PRENTICE, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat. Med.* **8** 431–440.
- ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90** 106–121. [MR1325118](#)
- SIBER, G. R., CHANG, I., BAKER, S., FERNSTEN, P., O’BRIEN, K. L. and SANTOSHAM, M. (2007). Estimating the protective concentration of anti-pneumococcal capsular polysaccharide antibodies. *Vaccine* **25** 3816–3826.
- STORSAETER, J., HALLANDER, H. O., GUSTAFSSON, L. and OLIN, P. (1998). Levels of anti-pertussis antibodies related to protection after household exposure to *Bordetella pertussis*. *Vaccine* **16** 1907–1916.
- VANDENBROUCKE, J. P. and PEARCE, N. (2012). Case–control studies: Basic concepts. *Int. J. Epidemiol.* **41** 1480–1489.

- VILLAR, L., DAYAN, G. H., ARREDONDO-GARCÍA, J. L., RIVERA, D. M., CUNHA, R. and DESEDA, C. (2015). Efficacy of a tetravalent dengue vaccine in children in Latin America. *N. Engl. J. Med.* **372** 113–123.
- WHITE, C. J., KUTER, B. J., NGAI, A., HILDEBRAND, C. S., ISGANITIS, K. L. and PATTERSON, C. M. (1992). Modified cases of chickenpox after varicella vaccination: Correlation of protection with antibody response. *Pediatr. Infect. Dis. J.* **11** 19–22.
- WORLD HEALTH ORGANIZATION (2017a). Dengue and severe dengue. Available at <http://www.who.int/mediacentre/factsheets/fs117/en/>. Accessed June, 2018.
- WORLD HEALTH ORGANIZATION (2017b). Dengue vaccine research. Available at http://www.who.int/immunization/research/development/dengue_vaccines/en/. Accessed June, 2018.

SEMIPARAMETRIC EMPIRICAL BEST PREDICTION FOR SMALL AREA ESTIMATION OF UNEMPLOYMENT INDICATORS

BY MARIA FRANCESCA MARINO^{*,1,4}, MARIA GIOVANNA RANALLI^{†,1},
NICOLA SALVATI^{‡,1,3} AND MARCO ALFÒ^{§,2}

University of Florence^{*}, *University of Perugia*[†], *University of Pisa*[‡] and
“*Sapienza*” *University of Rome*[§]

The Italian National Institute for Statistics regularly provides estimates of unemployment indicators using data from the labor force survey. However, direct estimates of unemployment incidence cannot be released for local labor market areas. These are unplanned domains defined as clusters of municipalities; many are out-of-sample areas, and the majority is characterized by a small sample size which renders direct estimates inadequate. The empirical best predictor represents an appropriate, model-based alternative. However, for non-Gaussian responses its computation and the computation of the analytic approximation to its mean squared error require the solution of (possibly) multiple integrals that, generally, have not a closed form. To solve the issue, Monte Carlo methods and parametric bootstrap are common choices, even though the computational burden is a nontrivial task. In this paper, we propose a semiparametric empirical best predictor for a (possibly) nonlinear mixed effect model by leaving the distribution of the area-specific random effects unspecified and estimating it from the observed data. This approach is known to lead to a discrete mixing distribution which helps avoid unverifiable parametric assumptions and heavy integral approximations. We also derive a second-order, bias-corrected analytic approximation to the corresponding mean squared error. Finite sample properties of the proposed approach are tested via a large scale simulation study. Furthermore, the proposal is applied to unit-level data from the 2012 Italian Labor Force Survey to estimate unemployment incidence for 611 local labor market areas using auxiliary information from administrative registers and the 2011 Census.

REFERENCES

- AITKIN, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Stat. Comput.* **6** 251–262.
- AITKIN, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55** 117–128. [MR1705676](#)
- BATES, D., MÄCHLER, M., BOLKER, B. and WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67** 1–48.
- BATTESE, G., HARTER, R. and FULLER, W. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.* **83** 28–36.
- BÖHNING, D. (1982). Convergence of Simar’s algorithm for finding the maximum likelihood estimate of a compound Poisson process. *Ann. Statist.* **10** 1006–1008. [MR0663451](#)

Key words and phrases. Binary data, Exponential Family, finite mixtures, general parameters, mixed logistic model, unit-level model.

- BOUBETA, M., LOMBARDÍA, M. J. and MORALES, D. (2016). Empirical best prediction under area-level Poisson mixed models. *TEST* **25** 548–569. [MR3531842](#)
- BOUBETA, M., LOMBARDÍA, M. J. and MORALES, D. (2017). Poisson mixed models for studying the poverty in small areas. *Comput. Statist. Data Anal.* **107** 32–47. [MR3575057](#)
- BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.
- BROWN, G., CHAMBERS, R., HEADY, P. and HEASMAN, D. (2001). Evaluation of small area estimation methods: an application to unemployment estimates from the UK LFS. In *Proc. Statistics Canada Symp. Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, Hull: Statistics Canada.
- CHEN, J. H. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** 221–233. [MR1331665](#)
- CHEN, S. X. and LIU, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statist. Sinica* **7** 875–892. [MR1488647](#)
- D’ALÒ, M., FALORSI, S. and SOLARI, F. (2017). Space-time unit-level EBLUP for large data sets. *J. Off. Stat.* **33** 61–77.
- D’ALÒ, M., DI CONSIGLIO, L., FALORSI, S., RANALLI, M. G. and SOLARI, F. (2012). Use of spatial information in small area models for unemployment rate estimation at sub-provincial areas in Italy. *J. Indian Soc. Agricultural Statist.* **66** 43–53. [MR2953458](#)
- 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](#)
- GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. [MR2847988](#)
- GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D. and SANTAMARÍA, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Comput. Statist. Data Anal.* **51** 2720–2733. [MR2338998](#)
- HOBZA, T. and MORALES, D. (2016). Empirical best prediction under unit-level logit mixed models. *J. Off. Stat.* **32** 661–692.
- HOBZA, T., MORALES, D. and SANTAMARÍA, L. (2018). Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *TEST* **27** 270–294. [MR3799140](#)
- JIANG, J. (1998). Consistent estimators in generalized linear mixed models. *J. Amer. Statist. Assoc.* **93** 720–729. [MR1631373](#)
- JIANG, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *J. Statist. Plann. Inference* **111** 117–127. [MR1955876](#)
- JIANG, J. and LAHIRI, P. (2001). Empirical best prediction for small area inference with binary data. *Ann. Inst. Statist. Math.* **53** 217–243. [MR1841133](#)
- LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* **73** 805–811. [MR0521328](#)
- LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.
- LINDSAY, B. G. (1983a). The geometry of mixture likelihoods: a general theory. *Ann. Statist.* **11** 86–94. [MR0684866](#)
- LINDSAY, B. G. (1983b). The geometry of mixture likelihoods. II. The Exponential Family. *Ann. Statist.* **11** 783–792. [MR0707929](#)
- LINDSAY, B. G. and LESPERANCE, M. L. (1995). A review of semiparametric mixture models. *J. Statist. Plann. Inference* **47** 29–39. [MR1360957](#)
- LÓPEZ-VIZCAÍNO, E., LOMBARDÍA, M. J. and MORALES, D. (2013). Multinomial-based small area estimation of labour force indicators. *Stat. Model.* **13** 153–178. [MR3179522](#)
- MARINO, M. F., RANALLI, M. G., SALVATI, N. and ALFÒ, M. (2019). Supplement to “Semi-Parametric Empirical Best Prediction for small area estimation of unemployment indicators.” DOI:10.1214/18-AOAS1226SUPP.

- MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92** 162–170. [MR1436105](#)
- MOLINA, I., NANDRAM, B. and RAO, J. N. K. (2014). Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach. *Ann. Appl. Stat.* **8** 852–885. [MR3262537](#)
- MOLINA, I., SAEI, A. and LOMBARDÍA, M. J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *J. Roy. Statist. Soc. Ser. A* **170** 975–1000. [MR2408988](#)
- MUNDLAK, Y. (1978). On the pooling of time series and cross section data. *Econometrica* **46** 69–85. [MR0478489](#)
- NEUHAUS, J. M. and MCCULLOCH, C. E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 859–872. [MR2301298](#)
- OAKES, D. (1999). Direct calculation of the information matrix via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 479–482. [MR1680298](#)
- PINHEIRO, J. and BATES, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J. Comput. Graph. Statist.* **4** 12–35.
- PRASAD, N. G. N. and RAO, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *J. Amer. Statist. Assoc.* **85** 163–171. [MR1137362](#)
- RAO, J. N. K. and MOLINA, I. (2015). *Small Area Estimation*, 2nd ed. *Wiley Series in Survey Methodology*. Wiley, Hoboken, NJ. [MR3380626](#)
- REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. [MR0738930](#)
- RODRIGUEZ, G. and GOLDMAN, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *J. Roy. Statist. Soc. Ser. A* **158** 73–89.
- SAEI, A. and CHAMBERS, R. (2003). Small area estimation under linear and generalized linear mixed models with time and area effects. In *S3RI Methodology Working Papers* 1–35. Southampton Statistical Sciences Research Institute, Southampton.
- SIMAR, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.* **4** 1200–1209. [MR0428582](#)
- VENABLES, W. N. and RIPLEY, B. D. (1994). *Modern Applied Statistics with S-Plus. Statistics and Computing*. Springer, New York. [MR1337030](#)

THE IDENTITY OF THE ZERO-TRUNCATED, ONE-INFLATED LIKELIHOOD AND THE ZERO-ONE-TRUNCATED LIKELIHOOD FOR GENERAL COUNT DENSITIES WITH AN APPLICATION TO DRINK-DRIVING IN BRITAIN

BY DANKMAR BÖHNING AND PETER G. M. VAN DER HEIJDEN

University of Southampton and University of Utrecht

For zero-truncated count data, as they typically arise in capture-recapture modelling, we consider modelling under one-inflation. This is motivated by police data on drink-driving in Britain which shows high one-inflation. The data, which are used here, are from the years 2011 to 2015 and are based on DR10 endorsements. We show that inference for an arbitrary count density with one-inflation can be equivalently based upon the associated zero-one truncated count density. This simplifies inference considerably including maximum likelihood estimation and likelihood ratio testing. For the drink-driving application, we use the geometric distribution which shows a good fit. We estimate the total drink-driving as about 2,300,000 drink-drivers in the observational period. As 227,578 were observed, this means that only about 10% of the drink-driving population is observed with a bootstrap confidence interval of 9%–12%.

REFERENCES

- ANAN, O., BÖHNING, D. and MARUOTTI, A. (2017). Uncertainty estimation in heterogeneous capture-recapture count data. *J. Stat. Comput. Simul.* **87** 2094–2114. [MR3647589](#)
- BÖHNING, D., BUNGE, J. and VAN DER HEIJDEN, P. G. M. (2018). Uncertainty assessment in capture-recapture studies and the choice of sampling effort. In *Capture-Recapture Methods for the Social and Medical Sciences* (D. Böhning, P. G. M. van der Heijden and J. Bunge, eds.) 389–394. CRC Press/CRC, Boca Raton, FL.
- BÖHNING, D., VAN DER HEIJDEN, P. G. M. and BUNGE, J. (2018). *Capture-Recapture Methods for the Social and Medical Sciences*. CRC Press/CRC, Boca Raton, FL.
- BORCHERS, D. L., BUCKLAND, S. T. and ZUCCHINI, W. (2004). *Estimating Animal Abundance. Closed Populations*. Springer, Heidelberg.
- BUNGE, J. and FITZPATRICK, M. (1993). Estimating the number of species: A review. *J. Amer. Statist. Assoc.* **88** 364–373.
- BUNGE, J., WILLIS, A. and WALSH, F. (2014). Estimating the number of species in microbial diversity studies. *Ann. Rev. Stat. Appl.* **1** 427–445.
- DEPARTMENT OF TRANSPORT (2018). Driving Licences. Available at <https://www.ethnicity-facts-figures.service.gov.uk/culture-and-community/transport/driving-licences/latest>, accessed 1/10/2018.
- EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. [MR1270903](#)

Key words and phrases. Capture–recapture, Chao estimator, behavioral response, power series distribution, mixture model, zero-truncated model, nonparametric estimator of population size.

- GODWIN, R. T. and BÖHNING, D. (2017). Estimation of the population size by using the one-inflated positive Poisson model. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 425–448. [MR3611695](#)
- KERIBIN, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A* **62** 49–66. [MR1769735](#)
- MCCREA, R. S. and MORGAN, B. J. T. (2015). *Analysis of Capture-Recapture Data*. Chapman & Hall/CRC Interdisciplinary Statistics Series. CRC Press, Boca Raton, FL. [MR3330977](#)
- PUIG, P. and KOKONENDJI, C. C. (2018). Non-parametric estimation of the number of zeros in truncated count distributions. *Scand. J. Stat.* **45** 347–365. [MR3803593](#)
- RAY, S. and LINDSAY, B. G. (2008). Model selection in high dimensions: A quadratic-risk-based approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 95–118. [MR2412633](#)
- ROAD ACCIDENTS AND SAFETY STATISTICS (2018). Statistics and data about reported accidents and casualties on public roads. Available at <https://www.gov.uk/government/collections/road-accidents-and-safety-statistics>. Accessed 1/10/2018.
- SANATHANAN, L. (1977). Estimating the size of a truncated sample. *J. Amer. Statist. Assoc.* **72** 669–672. [MR0458693](#)
- SELF, S. G. and LIANG, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82** 605–610. [MR0898365](#)
- THE GUARDIAN (2016). More than 8000 people caught drink-driving twice in five years. (accessed 23/5/2018), in print p. 11. Available at www.theguardian.com/society/2016/dec/30/more-than-8000-people-caught-drink-driving-twice-in-five-years.
- VAN DER HEIJDEN, P. G. M., BUSTAMI, R., CRUYFF, M. J. L. F., ENGBERSEN, G. and VAN HOUWELINGEN, H. C. (2003). Point and interval estimation of the population size using the truncated Poisson regression model. *Stat. Model.* **3** 305–322. [MR2012155](#)
- WANG, J.-P. Z. and LINDSAY, B. G. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *J. Amer. Statist. Assoc.* **100** 942–959. [MR2201021](#)
- WANG, J.-P. and LINDSAY, B. G. (2008). An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Stat. Methodol.* **5** 30–45. [MR2416938](#)

PHYLOGENY-BASED TUMOR SUBCLONE IDENTIFICATION USING A BAYESIAN FEATURE ALLOCATION MODEL¹

BY LI ZENG, JOSHUA L. WARREN AND HONGYU ZHAO

Yale University

Tumor cells acquire different genetic alterations during the course of evolution in cancer patients. As a result of competition and selection, only a few subgroups of cells with distinct genotypes survive. These subgroups of cells are often referred to as subclones. In recent years, many statistical and computational methods have been developed to identify tumor subclones, leading to biologically significant discoveries and shedding light on tumor progression, metastasis, drug resistance and other processes. However, most existing methods are either not able to infer the phylogenetic structure among subclones, or not able to incorporate copy number variations (CNV). In this article, we propose SIFA (tumor Subclone Identification by Feature Allocation), a Bayesian model which takes into account both CNV and tumor phylogeny structure to infer tumor subclones. We compare the performance of SIFA with two other commonly used methods using simulation studies with varying sequencing depth, evolutionary tree size, and tree complexity. SIFA consistently yields better results in terms of *Rand Index* and cellularity estimation accuracy. The usefulness of SIFA is also demonstrated through its application to whole genome sequencing (WGS) samples from four patients in a breast cancer study.

REFERENCES

- APARICIO, S. and CALDAS, C. (2013). The implications of clonal genome evolution for cancer medicine. *N. Engl. J. Med.* **368** 842–851.
- BROOKS, S. P. and GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7** 434–455. [MR1665662](#)
- BURRELL, R. A., MCGRANAHAN, N., BARTEK, J. and SWANTON, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501** 338–345.
- CAMPBELL, P. J., PLEASANCE, E. D., STEPHENS, P. J., DICKS, E., RANCE, R., GOODHEAD, I., FOLLOWS, G. A., GREEN, A. R., FUTREAL, P. A. and STRATTON, M. R. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl. Acad. Sci. USA* **105** 13081–13086.
- CAMPBELL, T. M., CASTRO, M. A., DE SANTIAGO, I., FLETCHER, M. N., HALIM, S., PRATHALINGAM, R., PONDER, B. A. and MEYER, K. B. (2016). FGFR2 risk SNPs confer breast cancer risk by augmenting oestrogen responsiveness. *Carcinogenesis* **37** 741–750.
- CARTER, S. L., CIBULSKIS, K., HELMAN, E., MCKENNA, A., SHEN, H., ZACK, T., LAIRD, P. W., ONOFRIO, R. C., WINCKLER, W., WEIR, B. A. et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30** 413–421.

Key words and phrases. Intra-tumor heterogeneity, latent feature allocation, model selection, tumor evolution.

- DESHWAR, A. G., VEMBU, S., YUNG, C. K., JANG, G. H., STEIN, L. and MORRIS, Q. (2015). PhylowGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16** 35.
- GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2014). *Bayesian Data Analysis, Vol. 2*. Chapman & Hall/CRC, Boca Raton, FL, USA.
- GERLINGER, M., ROWAN, A. J., HORSWELL, S., LARKIN, J., ENDESFELDER, D., GRONROOS, E., MARTINEZ, P., MATTHEWS, N., STEWART, A., TARPEY, P. et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366** 883–892.
- GERLINGER, M., HORSWELL, S., LARKIN, J., ROWAN, A. J., SALM, M. P., VARELA, I., FISHER, R., MCGRANAHAN, N., MATTHEWS, N., SANTOS, C. R. et al. (2014). Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46** 225–233.
- GEYER, C. J. (1991). Markov chain Monte Carlo maximum likelihood. Available at <http://www.stat.umn.edu/geyer/f05/8931/c.pdf>.
- GONZALEZ-PEREZ, A., PEREZ-LLAMAS, C., DEU-PONS, J., TAMBORERO, D., SCHROEDER, M. P., JENE-SANZ, A., SANTOS, A. and LOPEZ-BIGAS, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10** 1081–1082.
- GREAVES, M. and MALEY, C. C. (2012). Clonal evolution in cancer. *Nature* **481** 306–313.
- JIANG, Y., QIU, Y., MINN, A. J. and ZHANG, N. R. (2016). Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl. Acad. Sci. USA* **113** E5528–E5537.
- JIAO, W., VEMBU, S., DESHWAR, A. G., STEIN, L. and MORRIS, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinform.* **15** 35.
- KIMURA, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61** 893.
- KLAMBAUER, G., SCHWARZBAUER, K., MAYR, A., CLEVERT, D.-A., MITTERECKER, A., BODENHOFER, U. and HOCHREITER, S. (2012). cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* **40** e69.
- KRESO, A., O'BRIEN, C. A., VAN GALEN, P., GAN, O. I., NOTTA, F., BROWN, A. M., NG, K., MA, J., WIENHOLDS, E., DUNANT, C. et al. (2013). Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science* **339** 543–548.
- LEE, J., MÜLLER, P., GULUKOTA, K. and JI, Y. (2015). A Bayesian feature allocation model for tumor heterogeneity. *Ann. Appl. Stat.* **9** 621–639. [MR3371328](#)
- LEE, J., MÜLLER, P., SENGUPTA, S., GULUKOTA, K. and JI, Y. (2016). Bayesian inference for intratumour heterogeneity in mutations and copy number variation. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 547–563. [MR3522952](#)
- LI, B. and LI, J. Z. (2014). A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol.* **15** 473.
- MAGUIRE, S. L., LEONIDOU, A., WAI, P., MARCHIÒ, C., NG, C. K., SAPINO, A., SALOMON, A.-V., REIS-FILHO, J. S., WEIGELT, B. and NATRAJAN, R. C. (2015). SF3B1 mutations constitute a novel therapeutic target in breast cancer. *J. Pathol.* **235** 571–580.
- MARASS, F., MOULIERE, F., YUAN, K., ROSENFELD, N. and MARKOWETZ, F. (2016). A phylogenetic latent feature model for clonal deconvolution. *Ann. Appl. Stat.* **10** 2377–2404. [MR3592061](#)
- MILLER, C. A., WHITE, B. S., DEES, N. D., GRIFFITH, M., WELCH, J. S., GRIFFITH, O. L., VIJ, R., TOMASSON, M. H., GRAUBERT, T. A., WALTER, M. J. et al. (2014). SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.* **10** e1003665.
- NEAL, R. M. (2003). Slice sampling. *Ann. Statist.* **31** 705–767. With discussions and a rejoinder by the author. [MR1994729](#)

- NILSEN, G., LIESTØL, K., VAN LOO, P., VOLLAN, H. K. M., EIDE, M. B., RUEDA, O. M., CHIN, S.-F., RUSSELL, R., BAUMBUSCH, L. O., CALDAS, C. et al. (2012). Copynumber: Efficient algorithms for single-and multi-track copy number segmentation. *BMC Genomics* **13** 591.
- NOWELL, P. C. (1976). The clonal evolution of tumor cell populations. *Science* **194** 23–28.
- OJAMIES, P. N., KONTRO, M., EDGREN, H., ELLONEN, P., LAGSTROM, S., ALMUSA, H., MIETTINEN, T., ELDFORS, S., TAMBORERO, D., WENNERBERG, K. et al. (2016). Responses of AML patients to tailored drug regimens: Monitoring cancer subclones by ultra-deep resequencing. *Cancer Res.* **76** 2378–2378.
- PARISI, F., ARIYAN, S., NARAYAN, D., BACCHIOCCHI, A., HOYT, K., CHENG, E., XU, F., LI, P., HALABAN, R. and KLUGER, Y. (2011). Detecting copy number status and uncovering subclonal markers in heterogeneous tumor biopsies. *BMC Genomics* **12** 230.
- RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.
- ROBINSON, T. J., LIU, J. C., VIZEACOMAR, F., SUN, T., MACLEAN, N., EGAN, S. E., SCHIMMER, A. D., DATTI, A. and ZACKSENHAUS, E. (2013). RB1 status in triple negative breast cancer cells dictates response to radiation treatment and selective therapeutic drugs. *PLoS ONE* **8** e78641.
- ROTH, A., KHATTRA, J., YAP, D., WAN, A., LAKS, E., BIELE, J., HA, G., APARICIO, S., BOUCHARD-CÔTÉ, A. and SHAH, S. P. (2014). PyClone: Statistical inference of clonal population structure in cancer. *Nat. Methods* **11** 396–398.
- RUIZ, C., LENKIEWICZ, E., EVERS, L., HOLLEY, T., ROBESON, A., KIEFER, J., DEMEURE, M. J., HOLLINGSWORTH, M. A., SHEN, M., PRUNKARD, D. et al. (2011). Advancing a clinically relevant perspective of the clonal nature of cancer. *Proc. Natl. Acad. Sci. USA* **108** 12054–12059.
- SCHUH, A., BECQ, J., HUMPHRAY, S., ALEXA, A., BURNS, A., CLIFFORD, R., FELLER, S. M., GROCOCK, R., HENDERSON, S., KHREBTUKOVA, I. et al. (2012). Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* **120** 4191–4196.
- WANG, J., CAZZATO, E., LADEWIG, E., FRATTINI, V., ROSENBLOOM, D. I., ZAIRIS, S., ABATE, F., LIU, Z., ELLIOTT, O., SHIN, Y.-J. et al. (2016). Clonal evolution of glioblastoma under therapy. *Nat. Genet.* **48** 768–776.
- WATANABE, S. (2013). A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14** 867–897. [MR3049492](#)
- YACHIDA, S., JONES, S., BOZIC, I., ANTAL, T., LEARY, R., FU, B., KAMIYAMA, M., HRUBAN, R. H., ESHLEMAN, J. R., NOWAK, M. A. et al. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467** 1114–1117.
- YATES, L. R., GERSTUNG, M., KNAPPSKOG, S., DESMEDT, C., GUNDEM, G., VAN LOO, P., AAS, T., ALEXANDROV, L. B., LARSIMONT, D., DAVIES, H. et al. (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21** 751–759.
- YUAN, K., SAKOPARNIG, T., MARKOWETZ, F. and BEERENWINKEL, N. (2015). BitPhylogeny: A probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol.* **16** 36.
- ZARE, H., WANG, J., HU, A., WEBER, K., SMITH, J., NICKERSON, D., SONG, C., WITTEN, D., BLAU, C. A. and NOBLE, W. S. (2014). Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.* **10** e1003703.
- ZENG, L., WARREN, J. L. and ZHAO, H. (2019). Supplement to “Phylogeny-based tumor subclone identification using a Bayesian feature allocation model.” DOI:[10.1214/18-AOAS1223SUPP](#).

ESTIMATING POPULATION AVERAGE CAUSAL EFFECTS IN THE PRESENCE OF NON-OVERLAP: THE EFFECT OF NATURAL GAS COMPRESSOR STATION EXPOSURE ON CANCER MORTALITY¹

BY RACHEL C. NETHERY*, FABRIZIA MEALLI[†] AND FRANCESCA DOMINICI*

*Harvard T.H. Chan School of Public Health** and *University of Florence[†]*

Most causal inference studies rely on the assumption of overlap to estimate population or sample average causal effects. When data suffer from non-overlap, estimation of these estimands requires reliance on model specifications due to poor data support. All existing methods to address non-overlap, such as trimming or down-weighting data in regions of poor data support, change the estimand so that inference cannot be made on the sample or the underlying population. In environmental health research settings where study results are often intended to influence policy, population-level inference may be critical and changes in the estimand can diminish the impact of the study results, because estimates may not be representative of effects in the population of interest to policymakers. Researchers may be willing to make additional, minimal modeling assumptions in order to preserve the ability to estimate population average causal effects. We seek to make two contributions on this topic. First, we propose a flexible, data-driven definition of propensity score overlap and non-overlap regions. Second, we develop a novel Bayesian framework to estimate population average causal effects with minor model dependence and appropriately large uncertainties in the presence of non-overlap and causal effect heterogeneity. In this approach the tasks of estimating causal effects in the overlap and non-overlap regions are delegated to two distinct models suited to the degree of data support in each region. Tree ensembles are used to nonparametrically estimate individual causal effects in the overlap region, where the data can speak for themselves. In the non-overlap region where insufficient data support means reliance on model specification is necessary, individual causal effects are estimated by extrapolating trends from the overlap region via a spline model. The promising performance of our method is demonstrated in simulations. Finally, we utilize our method to perform a novel investigation of the causal effect of natural gas compressor station exposure on cancer outcomes. Code and data to implement the method and reproduce all simulations and analyses is available on Github (<https://github.com/rachelnethery/overlap>).

REFERENCES

- AUSTIN, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* **46** 399–424.

Key words and phrases. Overlap, propensity score, Bayesian additive regression trees, splines, natural gas, cancer mortality.

- BONATO, V., BALADANDAYUTHAPANI, V., BROOM, B. M., SULMAN, E. P., ALDAPE, K. D. and DO, K.-A. (2011). Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics* **27** 359–367.
- CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. [MR2758172](#)
- COLE, S. R. and HERNÁN, M. A. (2008). Constructing inverse probability weights for marginal structural models. *Am. J. Epidemiol.* **168** 656–664.
- CRISTIANINI, N. and SHAW-ETAYLOR, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge Univ. Press, Cambridge.
- CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. [MR2482144](#)
- D’AMOUR, A., DENG, P., FELLER, A., LEI, L. and SEKHON, J. (2017). Overlap in observational studies with high-dimensional covariates. Available at [arXiv:1711.02582](#).
- DI, Q., DAI, L., WANG, Y., ZANOBETTI, A., CHOIRAT, C., SCHWARTZ, J. D. and DOMINICI, F. (2017a). Association of short-term exposure to air pollution with mortality in older adults. *JAMA* **318** 2446–2456.
- DI, Q., WANG, Y., ZANOBETTI, A., WANG, Y., KOUTRAKIS, P., CHOIRAT, C., DOMINICI, F. and SCHWARTZ, J. D. (2017b). Air pollution and mortality in the medicare population. *N. Engl. J. Med.* **376** 2513–2522.
- FINKEL, M. L. (2016). Shale gas development and cancer incidence in southwest Pennsylvania. *Public Health* **141** 198–206.
- GOLDING, B. and WATSON, W. (1999). Possible mechanisms of carcinogenesis after exposure to benzene. *IARC Sci. Publ.* **150** 75–88.
- GUTMAN, R. and RUBIN, D. B. (2015). Estimation of causal effects of binary treatments in unconfounded studies. *Stat. Med.* **34** 3381–3398. [MR3412639](#)
- HAHN, P. R., MURRAY, J. and CARVALHO, C. M. (2018). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. Preprint. Available at [arXiv:1706.09523v2](#).
- HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. [MR2816546](#)
- HILL, J. and SU, Y.-S. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children’s cognitive outcomes. *Ann. Appl. Stat.* **7** 1386–1420. [MR3127952](#)
- HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* **15** 199–236.
- KASSOTIS, C. D., TILLITT, D. E., LIN, C.-H., MCELROY, J. A. and NAGEL, S. C. (2016). Endocrine-disrupting chemicals and oil and natural gas operations: Potential environmental contamination and recommendations to assess complex environmental mixtures. *Environ. Health Perspect.* **124** 256.
- KINDO, B. P., WANG, H. and PEÑA, E. A. (2016). Multinomial probit Bayesian additive regression trees. *Stat* **5** 119–131. [MR3484083](#)
- KING, G. and ZENG, L. (2005). The dangers of extreme counterfactuals. *Polit. Anal.* **14** 131–159.
- KLOCZKO, N. (2015). Summary on compressor stations and health impacts. <https://www.environmentalhealthproject.org/sites/default/files/assets/downloads/a-brief-review-of-compressor-stations-11.2015.pdf>.
- LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. [MR3803473](#)
- LI, F. and THOMAS, L. E. (2018). Addressing extreme propensity scores via the overlap weights. *Am. J. Epidemiol.* **kwy201**.
- LINERO, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *J. Amer. Statist. Assoc.* **113** 626–636. [MR3832214](#)

- LIU, Y., TRASKIN, M., LORCH, S. A., GEORGE, E. I. and SMALL, D. (2015). Ensemble of trees approaches to risk adjustment for evaluating a hospital's performance. *Health Care Manage. Sci.* **18** 58–66.
- MALTONI, C., CILIBERTI, A., COTTI, G., CONTI, B. and BELPOGGI, F. (1989). Benzene, an experimental multipotential carcinogen: Results of the long-term bioassays performed at the Bologna Institute of Oncology. *Environ. Health Perspect.* **82** 109.
- MCKENZIE, L. M., ALLSHOUSE, W. B., BYERS, T. E., BEDRICK, E. J., SERDAR, B. and ADGATE, J. L. (2017). Childhood hematologic cancer and residential proximity to oil and gas development. *PLoS ONE* **12** e0170423.
- MESSERSMITH, D., BROCKETT, D. and LOVELAND, C. (2015). Understanding natural gas compressor stations. *Penn State Extension*.
- MOKDAD, A. H., DWYER-LINDGREN, L., FITZMAURICE, C., STUBBS, R. W., BERTOZZI-VILLA, A., MOROZOFF, C., CHARARA, R., ALLEN, C., NAGHAVI, M. et al. (2017). Trends and patterns of disparities in cancer mortality among US counties, 1980–2014. *JAMA* **317** 388–406.
- NETHERY, R. C., MEALLI, F. and DOMINICI, F. (2019). Supplement to “Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality.” DOI:10.1214/18-AOAS1231SUPP.
- OAK RIDGE NATIONAL LABORATORY (2017). Natural gas compressor stations. https://hifld-dhs-gii.opendata.arcgis.com/datasets/fd7d62905d194eba87d2ee18d1a244b3_0. Accessed: 2018-03-29.
- PELLEGRITI, G., FRASCA, F., REGALBUTO, C., SQUATRITO, S. and VIGNERI, R. (2013). World-wide increasing incidence of thyroid cancer: Update on epidemiology and risk factors. *J. Cancer Epidemiol.* **2013**.
- PENNSYLVANIA DEPARTMENT OF ENVIRONMENTAL PROTECTION (2010). Southwestern Pennsylvania Marcellus Shale short-term ambient air sampling report. http://www.dep.state.pa.us/dep/deputate/airwaste/aq/aqm/docs/Marcellus_SW_11-01-10.pdf. Accessed: 2018-03-29.
- PETERSEN, M. L., PORTER, K. E., GRUBER, S., WANG, Y. and VAN DER LAAN, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* **21** 31–54. MR2867537
- R CORE TEAM (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- ROBERT WOOD JOHNSON FOUNDATION (2014). 2014 county health rankings and roadmaps. Prepared by Social Explorer. Accessed: 2018-03-27.
- ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974
- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.
- RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.
- SCHMIDHUBER, J. (2015). Deep learning in neural networks: An overview. *Neural Netw.* **61** 85–117.
- SOUTHWEST PENNSYLVANIA ENVIRONMENTAL HEALTH PROJECT (2015). Summary on compressor stations and health impacts. <http://www.environmentalhealthproject.org/files/Summary%20Compressor-station-emissions-and-health-impacts-02.24.2015.pdf>. Accessed 2018-03-27.
- 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
- US CENSUS BUREAU (2014). American community survey 2014 (5 year estimates). Prepared by Social Explorer. Accessed: 2018-03-27.

- US EPA (2018). Endocrine disruptor screening program (EDSP) estrogen receptor bioactivity. <https://www.epa.gov/endocrine-disruption/endocrine-disruptor-screening-program-edsp-estrogen-receptor-bioactivity#main-content>. Accessed: 2018-03-27.
- WANG, C., PARMIGIANI, G. and DOMINICI, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* **68** 661–686. [MR3055168](#)
- WANG, C., DOMINICI, F., PARMIGIANI, G. and ZIGLER, C. M. (2015). Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics* **71** 654–665. [MR3402601](#)
- WESTREICH, D. and COLE, S. R. (2010). Invited commentary: Positivity in practice. *Am. J. Epidemiol.* **171** 674–677.
- WESTREICH, D., LESSLER, J. and FUNK, M. J. (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J. Clin. Epidemiol.* **63** 826–833.
- WOLF EAGLE ENVIRONMENTAL (2009). Town of Dish, Texas ambient air monitoring analysis final report. https://townofdish.com/objects/DISH_-_final_report_revised.pdf. Accessed: 2018-05-08.
- WORLD TRADE CENTER HEALTH PROGRAM (2015). Minimum latency & types or categories of cancer. <https://www.cdc.gov/wtc/pdfs/WTCHP-Minimum-Cancer-Latency-PP-01062015.pdf>. Accessed: 2018-03-29.
- YANG, S. and DING, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika* **105** 487–493. [MR3804418](#)

SURVIVAL ANALYSIS OF DNA MUTATION MOTIFS WITH PENALIZED PROPORTIONAL HAZARDS

BY JEAN FENG^{*,1,6}, DAVID A. SHAW^{†,3,4,6}, VLADIMIR N. MININ^{‡,3,7},
NOAH SIMON^{*,2,7} AND FREDERICK A. MATSEN IV^{†,3,4,5,7}

University of Washington^{*}, *Fred Hutchinson Cancer Research Center*[†] and
University of California, Irvine[‡]

Antibodies, an essential part of our immune system, develop through an intricate process to bind a wide array of pathogens. This process involves randomly mutating DNA sequences encoding these antibodies to find variants with improved binding, though mutations are not distributed uniformly across sequence sites. Immunologists observe this nonuniformity to be consistent with “mutation motifs” which are short DNA subsequences that affect how likely a given site is to experience a mutation. Quantifying the effect of motifs on mutation rates is challenging. A large number of possible motifs makes this statistical problem high dimensional, while the unobserved history of the mutation process leads to a nontrivial missing data problem. We introduce an ℓ_1 -penalized proportional hazards model to infer mutation motifs and their effects. In order to estimate model parameters, our method uses a Monte Carlo EM algorithm to marginalize over the unknown ordering of mutations. We show that our method performs better on simulated data compared to current methods and leads to more parsimonious models. The application of proportional hazards to mutation processes is, to our knowledge, novel and formalizes the current methods in a statistical framework that can be easily extended to analyze the effect of other biological features on mutation rates.

REFERENCES

- AGGARWALA, V. and VOIGHT, B. F. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.* **48** 349–355.
- BECK, A. and TEBoulLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. [MR2486527](#)
- CAFFO, B. S., JANK, W. and JONES, G. L. (2005). Ascent-based Monte Carlo expectation-maximization. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 235–251. [MR2137323](#)
- CHAHWAN, R., EDELMANN, W., SCHARFF, M. D. and ROA, S. (2012). AIDing antibody diversity by error-prone mismatch repair. *Semin. Immunol.* **24** 293–300.
- COHEN, R. M., KLEINSTEIN, S. H. and LOUZOUN, Y. (2011). Somatic hypermutation targeting is influenced by location within the immunoglobulin V region. *Mol. Immunol.* **48** 1477–1483.
- COWELL, L. G. and KEPLER, T. B. (2000). The nucleotide-replacement spectrum under somatic hypermutation exhibits microsequence dependence that is strand-symmetric and distinct from that under germline mutation. *J. Immunol.* **164** 1971–1976.

Key words and phrases. Antibody maturation, survival analysis, Monte Carlo expectation-maximization, lasso, somatic hypermutation.

- CUI, A., DI NIRO, R., VANDER HEIDEN, J. A., BRIGGS, A. W., ADAMS, K., GILBERT, T., O'CONNOR, K. C., VIGNEAULT, F., SHLOMCHIK, M. J. et al. (2016). A model of somatic hypermutation targeting in mice based on high-throughput Ig sequencing data. *J. Immunol.* **197** 3566–3574.
- 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](#)
- DEZEURE, R., BÜHLMANN, P., MEIER, L. and MEINSHAUSEN, N. (2015). High-dimensional inference: Confidence intervals, p -values and R-software hdi. *Statist. Sci.* **30** 533–558. [MR3432840](#)
- DUNN-WALTERS, D. K., DOGAN, A., BOURSIER, L., MACDONALD, C. M. and SPENCER, J. (1998). Base-specific sequences that bias somatic hypermutation deduced by analysis of out-of-frame human IgVH genes. *J. Immunol.* **160** 2360–2364.
- ELHANATI, Y., SETHNA, Z., MARCOU, Q., CALLAN, C. G. JR, MORA, T. and WALCZAK, A. M. (2015). Inferring processes underlying B-cell repertoire diversity. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **370** 20140243.
- FENG, J., SHAW, D. A., MININ, V. N., SIMON, N. and MATSEN IV, F. A. (2019). Supplement to “Survival analysis of DNA mutation motifs with penalized proportional hazards.” DOI:[10.1214/18-AOAS1233SUPP](#).
- GOGGINS, W. B., FINKELSTEIN, D. M., SCHOENFELD, D. A. and ZASLAVSKY, A. M. (1998). A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics* **54** 1498–1507.
- GUPTA, N. T., VANDER HEIDEN, J. A., UDUMAN, M., GADALA-MARIA, D., YAARI, G. and KLEINSTEIN, S. H. (2015). Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31** 3356–3358.
- HAYNES, B. F., KELSOE, G., HARRISON, S. C. and KEPLER, T. B. (2012). B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. *Nat. Biotechnol.* **30** 423–433.
- HE, L., SOK, D., AZADNIA, P., HSUEH, J., LANDAIS, E., SIMEK, M., KOFF, W. C., POIGNARD, P., BURTON, D. R. et al. (2014). Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule bar-coding. *Sci. Rep.* **4** 6778.
- HERSHBERG, U., UDUMAN, M., SHLOMCHIK, M. J. and KLEINSTEIN, S. H. (2008). Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int. Immunol.* **20** 683–694.
- HESTERBERG, T., CHOI, N. H., MEIER, L. and FRALEY, C. (2008). Least angle and l_1 penalized regression: A review. *Stat. Surv.* **2** 61–93. [MR2520981](#)
- HOBOLTH, A. (2008). A Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. *J. Comput. Graph. Statist.* **17** 138–162. [MR2424799](#)
- HOEHN, K. B., LUNTER, G. and PYBUS, O. G. (2017). A phylogenetic codon substitution model for antibody lineages. *Genetics* **206** 417–427.
- HWANG, D. G. and GREEN, P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. USA* **101** 13994–14001.
- HWANG, J. K., WANG, C., DU, Z., MEYERS, R. M., KEPLER, T. B., NEUBERG, D., KWONG, P. D., MASCOLA, J. R., JOYCE, M. G. et al. (2017). Sequence intrinsic somatic mutation mechanisms contribute to affinity maturation of VRC01-class HIV-1 broadly neutralizing antibodies. *Proc. Natl. Acad. Sci. USA* **114** 8614–8619.
- KALBFLEISCH, J. D. and PRENTICE, R. L. (2011). *The Statistical Analysis of Failure Time Data. Wiley Series in Probability and Mathematical Statistics* **360**. Wiley, New York. [MR0570114](#)
- LEEB, H., PÖTSCHER, B. M. and EWALD, K. (2015). On various confidence intervals post-model-selection. *Statist. Sci.* **30** 216–227. [MR3353104](#)

- LEFRANC, M.-P. (2014). Immunoglobulins: 25 years of immunoinformatics and IMGT-ONTOLOGY. *Biomolecules* **4** 1102–1139.
- LEFRANC, M.-P., GIUDICELLI, V., GINESTOUX, C., BODMER, J., MÜLLER, W., BONTROP, R., LEMAITRE, M., MALIK, A., BARBIÉ, V. et al. (1999). IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.* **27** 209–212.
- LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44** 226–233. [MR0676213](#)
- MCCOY, C. O., BEDFORD, T., MININ, V. N., BRADLEY, P., ROBINS, H. and MATSEN, F. A. IV (2015). Quantifying evolutionary constraints on B-cell affinity maturation. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **370** 20140244.
- METHOT, S. P. and DI NOIA, J. M. (2017). Chapter two—Molecular mechanisms of somatic hypermutation and class switch recombination. In *Advances in Immunology* (F. W. Alt, ed.) **133** 37–87. Academic Press, San Diego, CA.
- NESTEROV, Y. (2013). Gradient methods for minimizing composite functions. *Math. Program.* **140** 125–161. [MR3071865](#)
- PHAM, P., BRANSTEITTER, R., PETRUSKA, J. and GOODMAN, M. F. (2003). Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424** 103–107.
- RALPH, D. K. and MATSEN IV, F. A. (2016a). Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLoS Comput. Biol.* **12** 1–25.
- RALPH, D. K. and MATSEN IV, F. A. (2016b). Likelihood-based inference of B cell clonal families. *PLoS Comput. Biol.* **12** e1005086.
- ROGOZIN, I. B. and DIAZ, M. (2004). Cutting edge: DGYW/WRCH is a better predictor of mutability at G: C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step Activation-Induced Cytidine Deaminase-triggered process. *J. Immunol.* **172** 3382–3384.
- ROGOZIN, I. B. and KOLCHANOV, N. A. (1992). Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochim. Biophys. Acta* **1171** 11–18.
- ROGOZIN, I. B., PAVLOV, Y. I., BEBENEK, K., MATSUDA, T. and KUNKEL, T. A. (2001). Somatic mutation hotspots correlate with DNA polymerase η error spectrum. *Nat. Immunol.* **2** 530–536.
- SCHATZ, D. G. and JI, Y. (2011). Recombination centres and the orchestration of V (D) J recombination. *Nat. Rev., Immunol.* **11** 251–263.
- SHENG, Z., SCHRAMM, C. A., KONG, R., NISC COMPARATIVE SEQUENCING PROGRAM, MULLIKIN, J. C., MASCOLA, J. R., KWONG, P. D. and SHAPIRO, L. (2017). Gene-specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic hypermutation. *Front. Immunol.* **8** 537.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- TIBSHIRANI, R. et al. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395.
- TONEGAWA, S. (1983). Somatic generation of antibody diversity. *Nature* **302** 575–581.
- UDUMAN, M., YAARI, G., HERSHBERG, U., STERN, J. A., SHLOMCHIK, M. J. and KLEINSTEIN, S. H. (2011). Detecting selection in immunoglobulin sequences. *Nucleic Acids Res.* **39** W499–W504.
- WEI, G. C. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.
- WIEHE, K., BRADLEY, T., RYAN MEYERHOFF, R., HART, C., WILLIAMS, W. B., EASTERHOFF, D., FAISON, W. J., KEPLER, T. B., SAUNDERS, K. O. et al. (2018). Functional relevance of improbable antibody mutations for HIV broadly neutralizing antibody development. *Cell Host Microbe* **23** 759–765.

- YAARI, G. and KLEINSTEIN, S. H. (2015). Practical guidelines for B-cell receptor repertoire sequencing analysis. *Gen. Med.* **7** 121.
- YAARI, G., UDUMAN, M. and KLEINSTEIN, S. H. (2012). Quantifying selection in high-throughput immunoglobulin sequencing data sets. *Nucleic Acids Res.* **40** e134.
- YAARI, G., VANDER HEIDEN, J. A., UDUMAN, M., GADALA-MARIA, D., GUPTA, N., STERN, J. N. H., O'CONNOR, K. C., HAFLER, D. A., LASERSON, U. et al. (2013). Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front. Immunol.* **4** 358.
- YAARI, G., BENICHO, J. I. C., VANDER HEIDEN, J. A., KLEINSTEIN, S. H. and LOUZOUN, Y. (2015). The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **370** 20140242.
- YEAP, L.-S., HWANG, J. K., DU, Z., MEYERS, R. M., MENG, F.-L., JAKUBAUSKAITÈ, A., LIU, M., MANI, V., NEUBERG, D. et al. (2015). Sequence-intrinsic mechanisms that target AID mutational outcomes on antibody genes. *Cell* **163** 1124–1137.
- ZHAO, S., SHOJAIE, A. and WITTEN, D. (2017). In defense of the indefensible: A very naive approach to high-dimensional inference. Preprint. Available at [arXiv:1705.05543](https://arxiv.org/abs/1705.05543).
- ZHOU, Q. and LIU, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* **20** 909–916.

EARLY DIAGNOSIS OF NEUROLOGICAL DISEASE USING PEAK DEGENERATION AGES OF MULTIPLE BIOMARKERS¹

BY FEI GAO, YUANJIA WANG AND DONGLIN ZENG

*University of Washington, Columbia University and
 University of North Carolina at Chapel Hill*

Neurological diseases are due to the loss of structure or function of neurons that eventually leads to cognitive deficit, neuropsychiatric symptoms, and impaired activities of daily living. Identifying sensitive and specific biological and clinical markers for early diagnosis allows recruiting patients into a clinical trial to test therapeutic intervention. However, many biomarker studies considered a single biomarker at one time that fails to provide precise prediction for disease age at onset. In this paper, we use longitudinally collected measurements from multiple biomarkers and measurement error-corrected clinical diagnosis ages to identify which biomarkers and what features of biomarker trajectories are useful for early diagnosis. Specifically, we assume that the subject-specific biomarker trajectories depend on unobserved states of underlying latent variables with the conditional mean follows a nonlinear sigmoid shape. We show that peak degeneration age of the biomarker trajectory is useful for early diagnosis. We propose an Expectation-Maximization (EM) algorithm to obtain the maximum likelihood estimates of all parameters and conduct extensive simulation studies to examine the performance of the proposed methods. Finally, we apply our methods to studies of Alzheimer's disease and Huntington's disease and identify a few important biomarkers that can be used for early diagnosis.

REFERENCES

- BATEMAN, R. J., XIONG, C., BENZINGER, T. L. S., FAGAN, A. M., GOATE, A., FOX, N. C., MARCUS, D. S., CAIRNS, N. J., XIE, X. et al. (2012). Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *N. Engl. J. Med.* **367** 795–804.
- BERNARDO, J. M. (1979). Expected information as expected utility. *Ann. Statist.* **7** 686–690. [MR0527503](#)
- DUBOIS, B., FELDMAN, H. H., JACOVA, C., DEKOSKY, S. T., BARBERGER-GATEAU, P., CUMMINGS, J., DELACOURTE, A., GALASKO, D., GAUTHIER, S. et al. (2007). Research criteria for the diagnosis of Alzheimer's disease: Revising the NINCDS-ADRDA criteria. *Lancet Neurol.* **6** 734–746.
- FJELL, A. M., WALHOVD, K. B., FENNEMA-NOTESTINE, C., MCEVOY, L. K., HAGLER, D. J., HOLLAND, D., BREWER, J. B. and DALE, A. M. (2009). One-year brain atrophy evident in healthy aging. *J. Neurosci.* **29** 15223–15231.
- GAO, F., WANG, Y. and ZENG, D. (2019). Supplement to “Early diagnosis of neurological disease using peak degeneration ages of multiple biomarkers.” DOI:[10.1214/18-AOAS1236SUPPA](#), DOI:[10.1214/18-AOAS1236SUPPB](#).

Key words and phrases. Alzheimer's disease, Huntington's disease, inflection point, measurement error, nonlinear mixed effects model, sigmoid function.

- GARCIA, T., MARDER, K. and WANG, Y. (2017). Statistical modeling of Huntington disease onset. *Handb. Clin. Neurol.* **144** 47–61.
- GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 243–268. [MR2325275](#)
- GOOD, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. Ser. B* **14** 107–114. [MR0077033](#)
- HUNTINGTON STUDY GROUP (1996). Unified Huntington's disease rating scale: Reliability and consistency. *Mov. Disord.* **11** 136–142.
- HALL, C. B., LIPTON, R. B., SLIWINSKI, M. and STEWART, W. F. (2000). A change point model for estimating the onset of cognitive decline in preclinical Alzheimer's disease. *Stat. Med.* **19** 1555–1566.
- HALL, C. B., YING, J., KUO, L., SLIWINSKI, M., BUSCHKE, H., KATZ, M. and LIPTON, R. B. (2001). Estimation of bivariate measurements having different change points, with application to cognitive ageing. *Stat. Med.* **20** 3695–3714.
- HALL, C. B., YING, J., KUO, L. and LIPTON, R. B. (2003). Bayesian and profile likelihood change point methods for modeling cognitive function over time. *Comput. Statist. Data Anal.* **42** 91–109. [MR1963357](#)
- HAMPEL, H., BÜRGER, K., TEIPEL, S. J., BOKDE, A. L. W., ZETTERBERG, H. and BLENNOW, K. (2008). Core candidate neurochemical and imaging biomarkers of Alzheimer's disease. *Alzheimer's Dement.* **4** 38–48.
- HOGAN, J. W. and LAIRD, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Stat. Med.* **16** 239–257.
- HOPFIELD, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **79** 2554–2558. [MR0652033](#)
- JACK, C. R., KNOPMAN, D. S., JAGUST, W. J., SHAW, L. M., AISEN, P. S., WEINER, M. W., PETERSEN, R. C. and TROJANOWSKI, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* **9** 119–128.
- JACQMIN-GADDA, H., COMMENGES, D. and DARTIGUES, J.-F. (2006). Random changepoint model for joint modeling of cognitive decline and dementia. *Biometrics* **62** 254–260. [MR2226580](#)
- JEDYNIAK, B. M., LANG, A., LIU, B., KATZ, E., ZHANG, Y., WYMAN, B. T., RAUNIG, D., JEDYNIAK, C. P., CAFFO, B. et al. (2012). A computational neurodegenerative disease progression score: Method and results with the Alzheimer's disease neuroimaging initiative cohort. *NeuroImage* **63** 1478–1486.
- KREMER, B., GOLDBERG, P., ANDREW, S. E., THEILMANN, J., TELENUS, H., ZEISLER, J., SQUITIERI, F., LIN, B., BASSETT, A. et al. (1994). A worldwide study of the Huntington's disease mutation: The sensitivity and specificity of measuring CAG repeats. *N. Engl. J. Med.* **330** 1401–1406.
- LITTLE, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *J. Amer. Statist. Assoc.* **90** 1112–1121. [MR1354029](#)
- LONG, J. D., PAULSEN, J. S., MARDER, K., ZHANG, Y., KIM, J.-I. and MILLS, J. A. (2014). Tracking motor impairments in the progression of Huntington's disease. *Mov. Disord.* **29** 311–319.
- MACDONALD, M. E., AMBROSE, C. M., DUYAO, M. P., MYERS, R. H., LIN, C., SRINIDHI, L., BARNES, G., TAYLOR, S. A., JAMES, M. et al. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72** 971–983.
- NELSON, P. T., ALAFUZOFF, I., BIGIO, E. H., BOURAS, C., BRAAK, H., CAIRNS, N. J., CASTELLANI, R. J., CRAIN, B. J., DAVIES, P. et al. (2012). Correlation of Alzheimer disease neuropathologic changes with cognitive status: A review of the literature. *J. Neuropathol. Exp. Neurol.* **71** 362–381.
- PAULSEN, J., LONG, J., ROSS, C., HARRINGTON, D., ERWIN, C., WILLIAMS, J., WESTERVELT, J., JOHNSON, H., AYLWARD, E. et al. (2014a). Prediction of manifest Huntington's dis-

- ease with clinical and imaging measures: A prospective observational study. *Lancet Neurol.* **13** 1193–1201.
- PAULSEN, J. S., LONG, J. D., JOHNSON, H. J., AYLWARD, E. H., ROSS, C. A., WILLIAMS, J. K., NANCE, M. A., ERWIN, C. J., WESTERVELT, H. J. et al. (2014b). Clinical and biomarker changes in premanifest Huntington disease show trial feasibility: A decade of the PREDICT-HD study. *Front. Aging Neurosci.* **6** 1–11.
- RIZOPOULOS, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press/CRC, New York.
- RUBINSZTEIN, D. C., LEGGO, J., COLES, R., ALMQVIST, E., BIANCALANA, V., CASSIMAN, J.-J., CHOTAI, K., CONNARTY, M., CRAUFURD, D. et al. (1996). Phenotypic characterization of individuals with 30–40 CAG repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36–39 repeats. *Am. J. Hum. Genet.* **59** 16–22.
- SAMTANI, M. N., FARNUM, M., LOBANOV, V., YANG, E., RAGHAVAN, N., DiBERNARDO, A. and NARAYAN, V. (2012). An improved model for disease progression in patients from the Alzheimer's disease neuroimaging initiative. *J. Clin. Pharmacol.* **52** 629–644.
- SHAW, L. M., VANDERSTICHELE, H., KNAPIK-CZAJKA, M., CLARK, C. M., AISEN, P. S., PETERSEN, R. C., BLENNOW, K., SOARES, H., SIMON, A. et al. (2009). Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Ann. Neurol.* **65** 403–413.
- SPERLING, R. A., AISEN, P. S., BECKETT, L. A., BENNETT, D. A., CRAFT, S., FAGAN, A. M., IWATSUBO, T., JACK, C. R., KAYE, J. et al. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement.* **7** 280–292.
- TSIATIS, A. A. and DAVIDIAN, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statist. Sinica* **14** 809–834. [MR2087974](#)

CORRECTION

STATISTICAL MODELING AND ANALYSIS OF TRACE ELEMENT CONCENTRATIONS IN FORENSIC GLASS EVIDENCE

BY KAREN D. H. PAN¹ AND KAREN KAFADAR²

University of Virginia

REFERENCES

- ASTM INTERNATIONAL (2016). ASTM E2927-16e1. Standard test method for determination of trace elements in soda-lime glass samples using laser ablation inductively coupled plasma mass spectrometry for forensic comparisons. Available at <https://doi.org/10.1520/E2927-16E01>.
- PAN, K. D. H. and KAFADAR, K. (2018). Statistical modeling and analysis of trace element concentrations in forensic glass evidence. *Ann. Appl. Stat.* **12** 788–814. [MR3834286](#)
- PARK, S. and CARRIQUIRY, A. (2018). Glass data description. Available at <https://github.com/CSAFE-ISU/AOAS-2018-glass-manuscript>.
- PARK, S. and CARRIQUIRY, A. (2019). Learning algorithms to evaluate forensic glass evidence. *Ann. Appl. Stat.* **13** 1068–1102.

The Annals of Applied Statistics

Next Issues

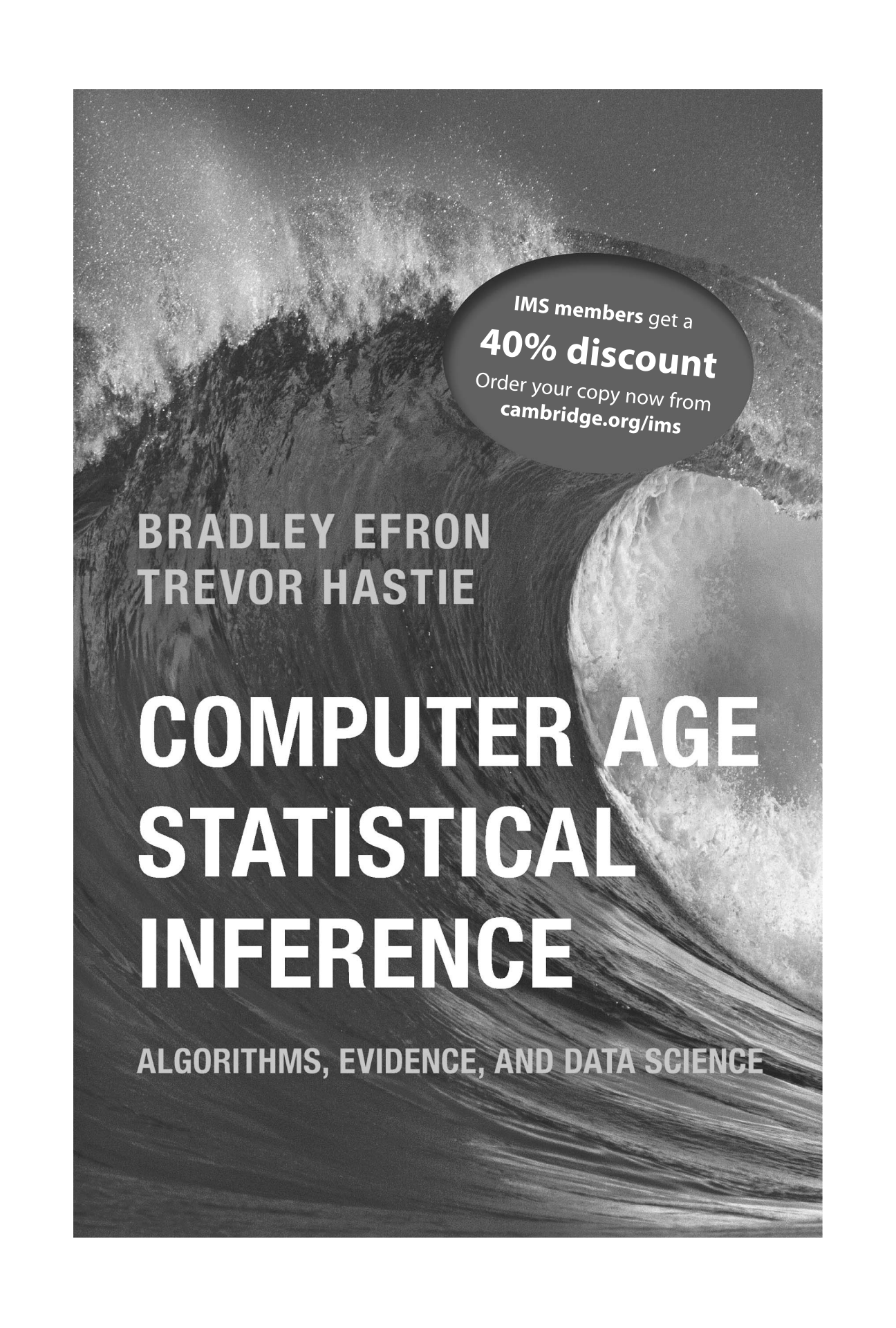
- Modeling seasonality and serial dependence of electricity price curves with warping functional autoregressive dynamics. YING CHEN, J. S. MARRON AND JIEJIE ZHANG
- Identifying and estimating principal causal effects in a multi-site trial of early college high schools LO-HUA YUAN, AVI FELLER AND LUKE MIRATRIX
- Radio-iBAG: Radiomics-based integrative Bayesian analysis of multiplatform genomic data
YOUYI ZHANG, JEFFREY MORRIS, SHIVALI NARANG AERRY,
ARVIND U. K. RAO AND VEERA BALADANDAYUTHAPANI
- Imputation and post-selection inference in models with missing data: An application to colorectal cancer surveillance guidelines
LIN LIU, YUQI QIU, LOKI NATARAJAN AND KAREN MESSER
- A hidden Markov model approach to characterizing the photo-switching behavior of fluorophores LEKHA PATEL, NILS GUSTAFSSON, YU LIN, RAIMUND OBER,
RICARDO HENRIQUES AND EDWARD COHEN
- The classification permutation test: A flexible approach to testing for covariate imbalance
JOHANN GAGNON-BARTSCH AND YOTAM SHEM-TOV
- Identifying multiple changes for a functional data sequence with application to freeway traffic segmentation. JENG-MIN CHIOU, YU-TING CHEN, TAILEN HSING
- Spatio-temporal short-term wind forecast: A calibrated regime-switching method
AHMED AZIZ EZZAT, MIKYOUNG JUN AND YU DING
- Network modelling of topological domains using Hi-C data. Y. X. RACHEL WANG,
PURNAMRITA SARKAR, OANA URSU, ANSHUL KUNDAJE AND PETER J. BICKEL
- Fast dynamic nonparametric distribution tracking in electron microscopic data
YANJUN QIAN, JIANHUA Z. HUANG, CHIWOO PARK AND YU DING
- Wavelet spectral testing: Application to nonstationary circadian rhythms
JESSICA KATE HARGREAVES, MARINA KNIGHT, JON PITCHFORD,
RACHAEL OAKENFULL, SANGEETA CHAWLA, JACK MUNNS AND SETH DAVIS
- Distributional regression forests for probabilistic precipitation forecasting in complex terrain
LISA SCHLOSSER, TORSTEN HOTHORN, RETO STAUFFER AND ACHIM ZEILEIS
- RCRnorm: An integrated system of random-coefficient hierarchical regression models for normalizing NanoString nCounter data GAOXIANG JIA, XINLEI WANG, QIWEI LI,
WEI LU, XIMING TANG, IGNACIO WISTUBA AND YANG XIE
- A hierarchical Bayesian model for single-cell clustering using RNA-sequencing data
YIYI LIU, JOSHUA WARREN AND HONGYU ZHAO
- Fitting a deeply-nested hierarchical model to a large book review dataset using a moment-based estimator NINGSHAN ZHANG, KYLE SCHMAUS AND PATRICK O. PERRY
- Network classification with applications to brain connectomics . . . JESUS D. ARROYO RELION,
DANIEL KESSLER, ELIZAVETA LEVINA AND STEPHAN F. TAYLOR
- Bayesian factor models for probabilistic cause of death assessment with verbal autopsies
TSUYOSHI KUNIHAMA, ZEHANG LI, SAMUEL CLARK AND
TYLER MCCORMICK
- A Bayesian mark interaction model for analysis of tumor pathology images
QIWEI LI, XINLEI WANG, FAMING LIANG AND GUANGHUA XIAO

Continued

The Annals of Applied Statistics

Next Issues—Continued

- Sequential decision model for inference and prediction on non-uniform hypergraphs with application to knot matching from computational forestry SEONG-HWAN JUN, SAMUEL W. K. WONG, JAMES V. ZIDEK AND ALEXANDRE BOUCHARD-CÔTÉ
- Incorporating conditional dependence in latent class models for probabilistic record linkage: Does it matter? HUIPING XU, XIAOCHUN LI, CHANGYU SHEN, SIU L. HUI AND SHAUN GRANNIS
- Bayesian modeling of the structural connectome for studying Alzheimer's disease
ARKAPRAVA ROY, SUBHASHIS GHOSAL, JEFFREY PRESCOTT AND
KINGSHUK ROY CHOUDHURY
- Approximate inference for constructing astronomical catalogs from images
JEFFREY REGIER, ANDREW MILLER, DAVID SCHLEGEL, RYAN ADAMS,
JON MCAULIFFE AND MR PRABHAT
- A latent discrete Markov random field approach to identifying and classifying historical forest communities based on spatial multivariate tree species counts. STEPHEN BERG, JUN ZHU, MURRAY K. CLAYTON, MONIKA E. SHEA AND DAVID J. MLADENOFF
- Bayesian methods for multiple mediators: Relating principal stratification and causal mediation in the analysis of power plant emission controls . . CHANMIN KIM, MICHAEL J. DANIELS, JOSEPH W. HOGAN, CHRISTINE CHOIRAT AND CORWIN M. ZIGLER
- Oblique random survival forests BYRON CASEY JAEGER, LEANN LONG, DUSTIN LONG, JEFF SZYCHOWSKI, YUAN-I MIN, LESLIE MCCLURE, GEORGE HOWARD AND
NOAH SIMON
- A nonparametric spatial test to identify factors that shape a microbiome . . SUSHEELA P. SINGH, ANA-MARIA STAIKU, ROBERT R. DUNN, NOAH FIERER AND BRIAN J. REICH
- Estimating the rate constant from biosensor data via an adaptive variational Bayesian approach YE ZHANG, ZHIGANG YAO, PATRIK FORSSEN AND TORGNY FORNSTEDT
- Estimating abundance from multiple sampling capture-recapture data via a multi-state multi-period stopover model HANNAH WORTHINGTON, RACHEL MCCREA, RUTH KING AND RICHARD GRIFFITHS
- SHOPPER: A probabilistic model of consumer choice with substitutes and complements
FRANCISCO J. R. RUIZ



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