

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

- A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data
YUNFAN TANG, LI MA AND DAN L. NICOLAE 1
- Phenomenological forecasting of disease incidence using heteroskedastic Gaussian
processes: A dengue case study LEAH R. JOHNSON, ROBERT B. GRAMACY,
JEREMY COHEN, ERIN MORDECAI, COURTNEY MURDOCK, JASON ROHR,
SADIE J. RYAN, ANNA M. STEWART-IBARRA AND DANIEL WEIKEL 27
- Spatial capture–recapture with partial identity: An application to camera traps
BEN C. AUGUSTINE, J. ANDREW ROYLE, MARCELLA J. KELLY,
CHRISTOPHER B. SATTER, ROBERT S. ALONSO, ERIN E. BOYDSTON
AND KEVIN R. CROOKS 67
- Simultaneous modelling of movement, measurement error, and observer dependence in
mark-recapture distance sampling: An application to Arctic bird surveys
PAUL B. CONN AND RAY T. ALISAUSKAS 96
- An empirical study of the maximal and total information coefficients and leading
measures of dependence DAVID N. RESHEF, YAKIR A. RESHEF,
PARDIS C. SABETI AND MICHAEL MITZENMACHER 123
- Covariate balancing propensity score for a continuous treatment: Application to the
efficacy of political advertisements CHRISTIAN FONG,
CHAD HAZLETT AND KOSUKE IMAI 156
- Integrative exploration of large high-dimensional datasets
CHRISTOPHER PARDY, SALLY GALBRAITH AND SUSAN R. WILSON 178
- Adjustment of nonconfounding covariates in case-control genetic association studies
HONG ZHANG, NILANJAN CHATTERJEE, DANIEL RADER AND JINBO CHEN 200
- How Gaussian mixture models might miss detecting factors that impact growth patterns
BRIANNA C. HEGGESETH AND NICHOLAS P. JEWELL 222
- Extreme value modelling of water-related insurance claims CHRISTIAN ROHRBECK,
EMMA F. EASTOE, ARNOLDO FRIGESSI AND JONATHAN A. TAWN 246
- Time-varying extreme value dependence with application to leading European stock
markets DANIELA CASTRO-CAMILO, MIGUEL DE CARVALHO
AND JENNIFER WADSWORTH 283
- Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests
with adjustment for false discovery rate BRIAN BADER, JUN YAN
AND XUEBIN ZHANG 310
- Design of vaccine trials during outbreaks with and without a delayed vaccination
comparator NATALIE E. DEAN, M. ELIZABETH HALLORAN
AND IRA M. LONGINI 330
- Two-level structural sparsity regularization for identifying lattices and defects in noisy
images XIN LI, ALEX BELIANINOV, ONDREJ DYCK, STEPHEN JESSE
AND CHIWOO PARK 348
- A spatio-temporal modeling framework for weather radar image data in tropical Southeast
Asia XIAO LIU, VIKNESWARAN GOPAL AND JAYANT KALAGNANAM 378

continued

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—*Continued from front cover*

- Multivariate integer-valued time series with flexible autocovariances and their application to major hurricane counts JAMES LIVSEY, ROBERT LUND, STEFANOS KECHAGIAS AND VLADAS PIPIRAS 408
- Stochastic simulation of predictive space–time scenarios of wind speed using observations and physical model outputs
JULIE BESSAC, EMIL CONSTANTINESCU AND MIHAI ANITESCU 432
- A multi-resolution model for non-Gaussian random fields on a sphere with application to ionospheric electrostatic potentials MINJIE FAN, DEBASHIS PAUL, THOMAS C. M. LEE AND TOMOKO MATSUO 459
- Reducing storage of global wind ensembles with stochastic generators
JAEHONG JEONG, STEFANO CASTRUCCIO, PAOLA CRIPPA AND MARC G. GENTON 490
- MSIQ: Joint modeling of multiple RNA-seq samples for accurate isoform quantification
WEI VIVIAN LI, ANQI ZHAO, SHIHUA ZHANG AND JINGYI JESSICA LI 510
- Kernel-penalized regression for analysis of microbiome data
TIMOTHY W. RANDOLPH, SEN ZHAO, WADE COPELAND, MEREDITH HULLAR AND ALI SHOJAIE 540
- Powerful test based on conditional effects for genome-wide screening
YAOWU LIU AND JUN XIE 567
- Fast inference of individual admixture coefficients using geographic data
KEVIN CAYE, FLORA JAY, OLIVIER MICHEL AND OLIVIER FRANÇOIS 586
- A unified statistical framework for single cell and bulk RNA sequencing data
LINGXUE ZHU, JING LEI, BERNIE DEVLIN AND KATHRYN ROEDER 609
- Modeling and estimation for self-exciting spatio-temporal models of terrorist activity
NICHOLAS J. CLARK AND PHILIP M. DIXON 633
- Empirical assessment of programs to promote collaboration: A network model approach
KATHERINE R. MCLAUGHLIN AND JOSHUA D. EMBREE 654

THE ANNALS OF APPLIED STATISTICS

Vol. 12, No. 1, pp. 1–682 March 2018

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:** Alison Etheridge, Department of Statistics, University of Oxford, Oxford, OX1 3LB, United Kingdom
- President-Elect:** Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA
- Past President:** Jon Wellner, Department of Statistics, University of Washington, Seattle, Washington 98195-4322, USA
- Executive Secretary:** Edsel Peña, Department of Statistics, University of South Carolina, Columbia, South Carolina 29208, USA
- Treasurer:** Zhengjun Zhang, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706-1510, USA
- Program Secretary:** Judith Rousseau, Université Paris Dauphine, Place du Maréchal DeLattre de Tassigny, 75016 Paris, France

IMS PUBLICATIONS

- The Annals of Statistics.** *Editors:* Edward I. George, Department of Statistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; Tailen Hsing, Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109-1107, USA
- The Annals of Applied Statistics.** *Editor-In-Chief:* Tilmann Gneiting, Heidelberg Institute for Theoretical Studies, HITS gGmbH, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany, and Institute for Stochastics, Karlsruhe Institute of Technology, Englerstr. 2, 76128 Karlsruhe, Germany
- 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.** *Editor:* Bálint Tóth, School of Mathematics, University of Bristol, University Walk, BS8 1TW, Bristol, U.K. and Institute of Mathematics, Technical University Budapest, Egry József u. 1, H-1111 Budapest, Hungary
- 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 12, Number 1, March 2018. 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.

A PHYLOGENETIC SCAN TEST ON A DIRICHLET-TREE MULTINOMIAL MODEL FOR MICROBIOME DATA

BY YUNFAN TANG*, LI MA^{†,1} AND DAN L. NICOLAE*,²

*University of Chicago** and *Duke University*[†]

In this paper, we introduce the phylogenetic scan test (PhyloScan) for investigating cross-group differences in microbiome compositions using the Dirichlet-tree multinomial (DTM) model. DTM models the microbiome data through a cascade of independent local DMs on the internal nodes of the phylogenetic tree. Each of the local DMs captures the count distributions of a certain number of operational taxonomic units at a given resolution. Since distributional differences tend to occur in clusters along evolutionary lineages, we design a scan statistic over the phylogenetic tree to allow nodes to borrow signal strength from their parents and children. We also derive a formula to bound the tail probability of the scan statistic, and verify its accuracy through simulations. The PhyloScan procedure is applied to the American Gut dataset to identify taxa associated with diet habits. Empirical studies performed on this dataset show that PhyloScan achieves higher testing power in most cases.

REFERENCES

- CAPORASO, J. G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F. D., COSTELLO, E. K., FIERER, N., PENA, A. G., GOODRICH, J. K., GORDON, J. I. et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7** 335–336.
- CHEN, Y. and HANSON, T. E. (2014). Bayesian nonparametric k -sample tests for censored and uncensored data. *Comput. Statist. Data Anal.* **71** 335–346. [MR3131974](#)
- CHEN, J. and LI, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* **7** 418–442. [MR3086425](#)
- DAVID, L. A., MAURICE, C. F., CARMODY, R. C., GOOTENBERG, D. B., BUTTON, J. E., WOLFE, B. E., LING, A. V., DEVLIN, A. S., VARMA, Y., FISCHBACH, M. A. et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505** 559–563.
- DENNIS, S. Y. III (1991). On the hyper-Dirichlet type 1 and hyper-Liouville distributions. *Comm. Statist. Theory Methods* **20** 4069–4081. [MR1158563](#)
- DOHMEN, K. (2000). Improved Bonferroni inequalities via union-closed set systems. *J. Combin. Theory Ser. A* **92** 61–67. [MR1783939](#)
- DOHMEN, K. (2002). Improved inclusion-exclusion identities and Bonferroni inequalities with reliability applications. *SIAM J. Discrete Math.* **16** 156–171. [MR1972081](#)
- DOHMEN, K. and TITTMANN, P. (2004). Bonferroni–Galambos inequalities for partition lattices. *Electron. J. Combin.* **11** Article ID 85. [MR2114189](#)
- EFRON, B. (1997). The length heuristic for simultaneous hypothesis tests. *Biometrika* **84** 143–157. [MR1450198](#)
- GLAZ, J., NAUS, J. and WALLENSTEIN, S. (2001). *Scan Statistics*. Springer, New York. [MR1869112](#)

Key words and phrases. Dirichlet-multinomial, microbiome, Dirichlet-tree multinomial, phylogenetic tree, PhyloScan, scan statistics, union probability.

- HAHN, T. (2005). Cuba—A library for multidimensional numerical integration. *Comput. Phys. Commun.* **168** 78–95. [MR2136794](#)
- HOLMES, I., HARRIS, K. and QUINCE, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS ONE* **7** Article ID e30126.
- HOLMES, C. C., CARON, F., GRIFFIN, J. E. and STEPHENS, D. A. (2015). Two-sample Bayesian nonparametric hypothesis testing. *Bayesian Anal.* **10** 297–320. [MR3420884](#)
- HUMAN MICROBIOME PROJECT CONSORTIUM (2012). A framework for human microbiome research. *Nature* **486** 215–221.
- HUNTER, D. (1976). An upper bound for the probability of a union. *J. Appl. Probab.* **13** 597–603. [MR0415722](#)
- LAVINE, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **20** 1222–1235. [MR1186248](#)
- LA ROSA, P. S., BROOKS, J. P., DEYCH, E., BOONE, E. L., EDWARDS, D. J., WANG, Q., SODERGREN, E., WEINSTOCK, G. and SHANNON, W. D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE* **7** Article ID e52078. DOI:[10.1371/journal.pone.0052078](#).
- MA, L. and WONG, W. H. (2011). Coupling optional Pólya trees and the two sample problem. *J. Amer. Statist. Assoc.* **106** 1553–1565. [MR2896856](#)
- MCDONALD, D., BIRMINGHAM, A. and KNIGHT, R. (2015a). Context and the human microbiome. *Microbiome* **3** 1–8.
- MCDONALD, D., HORNIG, M., LOZUPONE, C., DEBELIUS, J., GILBERT, J. and KNIGHT, R. (2015b). Towards large-cohort comparative studies to define the factors influencing the gut microbial community structure of ASD patients. *Microb. Ecol. Health Dis.* **26** 26555.
- MOSIMANN, J. E. (1962). On the compound multinomial distribution, the multivariate β -distribution, and correlations among proportions. *Biometrika* **49** 65–82. [MR0143299](#)
- NAIMAN, D. Q. and WYNN, H. P. (1992). Inclusion-exclusion-Bonferroni identities and inequalities for discrete tube-like problems via Euler characteristics. *Ann. Statist.* **20** 43–76. [MR1150334](#)
- NAIMAN, D. Q. and WYNN, H. P. (1997). Abstract tubes, improved inclusion-exclusion identities and inequalities and importance sampling. *Ann. Statist.* **25** 1954–1983. [MR1474076](#)
- NEUMAN, H., DEBELIUS, J. W., KNIGHT, R. and KOREN, O. (2015). Microbial endocrinology: The interplay between the microbiota and the endocrine system. *FEMS Microbiol. Rev.* **39** 509–521.
- SILVERMAN, J. D., WASHBURNE, A., MUKHERJEE, S. and DAVID, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *ELife* **6** Article ID e21887.
- SORIANO, J. and MA, L. (2017). Probabilistic multi-resolution scanning for two-sample differences. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 547–572. [MR3611759](#)
- TANG, Y., MA, L. and NICOLAE, D. L. (2018). Supplement to “A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data.” DOI:[10.1214/17-AOAS1086SUPP](#).
- TANG, Z., CHEN, G., ALEKSEYENKO, A. V. and LI, H. (2017). A general framework for association analysis of microbial communities on a taxonomic tree. *Bioinformatics* **33** 1278–1285. DOI:[10.1093/bioinformatics/btw804](#).
- TAYLOR, J. E., WORSLEY, K. J. and GOSSELIN, F. (2007). Maxima of discretely sampled random fields, with an application to ‘bubbles’. *Biometrika* **94** 1–18. [MR2307898](#)
- TURNBAUGH, P. J., RIDAURA, V. K., FAITH, J. J., REY, F. E., KNIGHT, R. and GORDON, J. I. (2014). The effect of diet on the human gut microbiome: A metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1** Article ID 6ra14. DOI:[10.1126/scitranslmed.3000322](#).
- WANG, T. and ZHAO, H. (2017). A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics* **73** 792–801.
- WEIR, B. S. and HILL, W. G. (2002). Estimating F-statistics. *Annu. Rev. Genet.* **36** 721–750.
- WORSLEY, K. J. (1982). An improved Bonferroni inequality and applications. *Biometrika* **69** 297–302. [MR0671966](#)

PHENOMENOLOGICAL FORECASTING OF DISEASE INCIDENCE USING HETEROSKEDASTIC GAUSSIAN PROCESSES: A DENGUE CASE STUDY

BY LEAH R. JOHNSON^{*,1}, ROBERT B. GRAMACY^{*,2}, JEREMY COHEN[†],
ERIN MORDECAI^{‡,1,3}, COURTNEY MURDOCK[§], JASON ROHR^{†,1,4},
SADIE J. RYAN^{¶,1}, ANNA M. STEWART-IBARRA^{||,1} AND DANIEL WEIKEL^{**}

Virginia Tech^{*}, *University of South Florida*[†], *Stanford University*[‡],
University of Georgia[§], *University of Florida*[¶],
SUNY Upstate Medical University^{||} and *University of Michigan*^{**}

In 2015 the US federal government sponsored a dengue forecasting competition using historical case data from Iquitos, Peru and San Juan, Puerto Rico. Competitors were evaluated on several aspects of out-of-sample forecasts including the *targets* of peak week, peak incidence during that week, and total season incidence across each of several seasons. Our team was one of the winners of that competition, outperforming other teams in multiple targets/locales. In this paper we report on our methodology, a large component of which, surprisingly, ignores the known biology of epidemics at large—for example, relationships between dengue transmission and environmental factors—and instead relies on flexible nonparametric nonlinear Gaussian process (GP) regression fits that “memorize” the trajectories of past seasons, and then “match” the dynamics of the unfolding season to past ones in real-time. Our phenomenological approach has advantages in situations where disease dynamics are less well understood, or where measurements and forecasts of ancillary covariates like precipitation are unavailable, and/or where the strength of association with cases are as yet unknown. In particular, we show that the GP approach generally outperforms a more classical generalized linear (autoregressive) model (GLM) that we developed to utilize abundant covariate information. We illustrate variations of our method(s) on the two benchmark locales alongside a full summary of results submitted by other contest competitors.

REFERENCES

- ANKENMAN, B., NELSON, B. L. and STAUM, J. (2010). Stochastic kriging for simulation meta-modeling. *Oper. Res.* **58** 371–382. [MR2674803](#)
- BARRERA, R., AMADOR, M. and MACKAY, A. J. (2011). Population dynamics of *Aedes aegypti* and dengue as influenced by weather and human behavior in San Juan, Puerto Rico. *PLoS Negl. Trop. Dis.* **5** e1378.
- BINOIS, M., GRAMACY, R. B. and LUDKOVSKI, M. (2016). Practical heteroskedastic Gaussian process modeling for large simulation experiments. arXiv preprint, [arXiv:1611.05902](#).

Key words and phrases. Epidemiology, Gaussian process, heteroskedastic modeling, latent variable, generalized linear (autoregressive) model, dengue fever.

- BORNN, L., SHADDICK, G. and ZIDEK, J. V. (2012). Modeling nonstationary processes through dimension expansion. *J. Amer. Statist. Assoc.* **107** 281–289. [MR2949359](#)
- CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York. Revised reprint of the 1991 edition. [MR1239641](#)
- DEGALLIER, N., FAVIER, C., MENKES, C., LENGAINNE, M., RAMALHO, W. M., SOUZA, R., SERVAIN, J. and BOULANGER, J.-P. (2010). Toward an early warning system for dengue prevention: Modeling climate impact on dengue transmission. *Clim. Change* **98** 581–592.
- ELDERD, B. D., DUKIC, V. M. and DWYER, G. (2006). Uncertainty in predictions of disease spread and public health responses to bioterrorism and emerging diseases. *Proc. Natl. Acad. Sci. USA* **103** 15693–15697.
- FARAH, M., BIRRELL, P., CONTI, S. and DE ANGELIS, D. (2014). Bayesian emulation and calibration of a dynamic epidemic model for A/H1N1 influenza. *J. Amer. Statist. Assoc.* **109** 1398–1411. [MR3293599](#)
- GAGNON, A. S., BUSH, A. B. and SMOYER-TOMIC, K. E. (2001). Dengue epidemics and the El Niño southern oscillation. *Clim. Res.* **19** 35–43.
- GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. [MR2847988](#)
- GNEITING, T. (2017). When is the mode functional the Bayes classifier? *Stat* **6** 204–206. [MR3671158](#)
- GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. [MR2345548](#)
- GNEITING, T., LARSON, K., WESTRICK, K., GENTON, M. G. and ALDRICH, E. (2006). Calibrated probabilistic forecasting at the Stateline wind energy center: The regime-switching space-time method. *J. Amer. Statist. Assoc.* **101** 968–979. [MR2324108](#)
- GRAMACY, R. B. (2014). laGP: Local approximate Gaussian process regression. R package version 1.1-4.
- GRAMACY, R. B. (2016). laGP: Large-scale spatial modeling via local approximate Gaussian processes in R. *J. Stat. Softw.* **72** 1–46.
- HU, R. and LUDKOVSK, M. (2017). Sequential design for ranking response surfaces. *SIAM/ASA J. Uncertain. Quantificat.* **5** 212–239. [MR3614687](#)
- JOHANSSON, M. A., CUMMINGS, D. A. T. and GLASS, G. E. (2009). Multiyear climate variability and dengue–El Niño southern oscillation, weather, and dengue incidence in Puerto Rico, Mexico, and Thailand: A longitudinal data analysis. *PLoS Med.* **6** e1000168.
- JOHNSON, L. R. and GRAMACY, R. B. (2017). vbdcast: Vector-borne disease forecasting. Technical report. <https://github.com/lrjohnson0/vbdcast>.
- JOHNSON, L. R., BEN-HORIN, T., LAFFERTY, K. D., McNALLY, A., MORDECAI, E., PAAIJMANS, K. P., PAWAR, S. and RYAN, S. J. (2015). Understanding uncertainty in temperature effects on vector-borne disease: A Bayesian approach. *Ecology* **96** 203–213.
- JOHNSON, L. R., GRAMACY, R. B., COHEN, J., MORDECAI, E., MURDOCK, C., ROHR, J., RYAN, S. J., STEWART-IBARRA, A. M. and WEIKEL, D. (2018). Supplement to “Phenomenological forecasting of disease incidence using heteroskedastic Gaussian processes: A dengue case study.” DOI:10.1214/17-AOAS1090SUPPA, DOI:10.1214/17-AOAS1090SUPPB.
- KOEPKE, A. A., LONGINI JR., I. M., HALLORAN, M. E., WAKEFIELD, J. and MININ, V. N. (2016). Predictive modeling of cholera outbreaks in Bangladesh. *Ann. Appl. Stat.* **10** 575.
- KUHN, K., CAMPBELL-LENDRUM, D., HAINES, A., COX, J., CORVALÁN, C., ANKER, M. et al. (2005). Using climate to predict infectious disease epidemics. White Paper, World Health Organization, Geneva. www.who.int/globalchange/publications/infectdiseases/en/index.html.
- LAMBRECHTS, L., PAAIJMANS, K. P., FANSIRI, T., CARRINGTON, L. B., KRAMER, L. D., THOMAS, M. B. and SCOTT, T. W. (2011). Impact of daily temperature fluctuations on dengue virus transmission by *Aedes aegypti*. *Proc. Natl. Acad. Sci. USA* **108** 7460–7465.

- LUDKOVSKI, M. and NIEMI, J. (2010). Optimal dynamic policies for influenza management. *Stat. Commun. Infect. Dis.* **2** Art. 5, 27. [MR2764286](#)
- MATHERON, G. (1963). Principles of geostatistics. *Econ. Geol.* **58** 1246–1266.
- MERL, D., JOHNSON, L. R., GRAMACY, R. B. and MANGEL, M. (2009). A statistical framework for the adaptive management of epidemiological interventions. *PLoS ONE* **4** e5807.
- MOORE, C. G., CLINE, B. L., RUIZ-TIBÉN, E., LEE, D., ROMNEY-JOSEPH, H. and RIVERA-CORREA, E. (1978). *Aedes aegypti* in Puerto Rico: Environmental determinants of larval abundance and relation to dengue virus transmission. *Am. J. Trop. Med. Hyg.* **27** 1225–1231.
- MORDECAI, E. A., PAAIJMANS, K. P., JOHNSON, L. R., BALZER, C., BEN-HORIN, T., DE MOOR, E., McNALLY, A., PAWAR, S., RYAN, S. J., SMITH, T. C. and LAFFERTY, K. D. (2013). Optimal temperature for malaria transmission is dramatically lower than previously predicted. *Ecol. Lett.* **16** 22–30.
- MORDECAI, E., COHEN, J., EVANS, M. V., GUDAPATI, P., JOHNSON, L. R., LIPPI, C. A., MI-AZGOWICZ, K., MURDOCK, C. C., ROHR, J. R., RYAN, S. J., SAVAGE, V., SHOCKET, M., STEWART IBARRA, A., THOMAS, M. B. and WEIKEL, D. P. (2017). Detecting the impact of temperature on transmission of Zika, dengue and chikungunya using mechanistic models. *PLoS Negl. Trop. Dis.* **11** e0005568.
- OSTHUS, D., HICKMANN, K. S., CARAGEA, P. C., HIGDON, D. and DEL VALLE, S. Y. (2017). Forecasting seasonal influenza with a state–space SIR model. *Ann. Appl. Stat.* **11** 202–224. [MR3634321](#)
- R DEVELOPMENT CORE TEAM (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.
- RAY, E. L., SAKREJDA, K., LAUER, S. A., JOHANSSON, M. A. and REICH, N. G. (2017). Infectious disease prediction with kernel conditional density estimation. Technical report. github.com/reichlab/article-disease-pred-with-kcde.
- REYNOLDS, R. W., RAYNER, N. A., SMITH, T. M., STOKES, D. C. and WANG, W. (2002). An improved in situ and satellite SST analysis for climate. *J. Climate* **15** 1609–1625.
- SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. With comments and a rejoinder by the authors. [MR1041765](#)
- STEWART-IBARRA, A. M. and LOWE, R. (2013). Climate and non-climate drivers of dengue epidemics in southern coastal Ecuador. *Am. J. Trop. Med. Hyg.* **88** 971–981.
- STEWART-IBARRA, A. M., RYAN, S. J., BELTRÁN, E., MEJÍA, R., SILVA, M. and MUÑOZ, Á. (2013). Dengue vector dynamics (*Aedes aegypti*) influenced by climate and social factors in Ecuador: Implications for targeted control. *PLoS ONE* **8** e78263.
- THOMSON, M. C., GARCIA-HERRERA, R. and BENISTON, M. (2008). *Seasonal Forecasts, Climatic Change and Human Health*. Springer.
- VENABLES, W. N. and RIPLEY, B. D. (1994). *Modern Applied Statistics with S-Plus*. Springer, New York. [MR1337030](#)
- WORLD HEALTH ORGANIZATION (2009). Dengue: Guidelines for diagnosis, treatment, prevention and control. Special Programme for Research and Training in Tropical Diseases, Department of Control of Neglected Tropical Diseases, and Epidemic and Pandemic Alert, World Health Organization.
- WORLD HEALTH ORGANIZATION (2016). Dengue vaccine: WHO position paper—July 2016. *Weekly Epidemiological Record* **91** 349–364.
- XU, L., STIGE, L. C., CHAN, K.-S., ZHOU, J., YANG, J., SANG, S., WANG, M., YANG, Z., YAN, Z., JIANG, T., LU, L., YUE, Y., LIU, X., LIN, H., XU, J., LIU, Q. and STENSETH, N. C. (2016). Climate variation drives dengue dynamics. *Proc. Natl. Acad. Sci. USA* 201618558.

YAMANA, T. K., KANDULA, S. and SHAMAN, J. (2016). Superensemble forecasts of dengue outbreaks. *J. R. Soc. Interface* **13** 20160410.

SPATIAL CAPTURE–RECAPTURE WITH PARTIAL IDENTITY: AN APPLICATION TO CAMERA TRAPS

BY BEN C. AUGUSTINE*, J. ANDREW ROYLE[†], MARCELLA J. KELLY*,
 CHRISTOPHER B. SATTER*, ROBERT S. ALONSO*, ERIN E. BOYDSTON[‡] AND
 KEVIN R. CROOKS[§]

*Virginia Tech**, *U.S. Geological Survey, Patuxent Wildlife Research Center[†]*, *U.S.
 Geological Survey, Western Ecological Research Center[‡]* and *U.S. Geological
 Survey, Colorado State University[§]*

Camera trapping surveys frequently capture individuals whose identity is only known from a single flank. The most widely used methods for incorporating these partial identity individuals into density analyses discard some of the partial identity capture histories, reducing precision, and, while not previously recognized, introducing bias. Here, we present the spatial partial identity model (SPIM), which uses the spatial location where partial identity samples are captured to probabilistically resolve their complete identities, allowing all partial identity samples to be used in the analysis. We show that the SPIM outperforms other analytical alternatives. We then apply the SPIM to an ocelot data set collected on a trapping array with double-camera stations and a bobcat data set collected on a trapping array with single-camera stations. The SPIM improves inference in both cases and, in the ocelot example, individual sex is determined from photographs used to further resolve partial identities—one of which is resolved to near certainty. The SPIM opens the door for the investigation of trapping designs that deviate from the standard two camera design, the combination of other data types between which identities cannot be deterministically linked, and can be extended to the problem of partial genotypes.

REFERENCES

- ALONSO, R. S., MCCLINTOCK, B. T., LYREN, L. M., BOYDSTON, E. E. and CROOKS, K. R. (2015). Mark–recapture and Mark–resight methods for estimating abundance with remote cameras: A carnivore case study. *PLoS ONE* **10** e0123032.
- AUGUSTINE, B. C., ROYLE, J. A., KELLY, M. J., SATTER, C. B., ALONSO, R. S., BOYDSTON, E. E. and CROOKS, K. R. (2018a). Supplement to “Spatial capture–recapture with partial identity: An application to camera traps.” DOI:[10.1214/17-AOAS1091SUPPA](https://doi.org/10.1214/17-AOAS1091SUPPA).
- AUGUSTINE, B. C., ROYLE, J. A., KELLY, M. J., SATTER, C. B., ALONSO, R. S., BOYDSTON, E. E. and CROOKS, K. R. (2018b). Supplement to “Spatial capture–recapture with partial identity: An application to camera traps.” DOI:[10.1214/17-AOAS1091SUPPB](https://doi.org/10.1214/17-AOAS1091SUPPB).
- BONNER, S. J. and HOLMBERG, J. (2013). mark–recapture with multiple, non-invasive marks. *Bio-metrics* **69** 766–775. [MR3106605](https://doi.org/10.1111/biom.12060)

Key words and phrases. Spatial capture–recapture, partial identity, camera trapping, multiple marks.

- BORCHERS, D. L. and EFFORD, M. (2008). Spatially explicit maximum likelihood methods for capture–recapture studies. *Biometrics* **64** 377–385.
- CHANDLER, R. B. and CLARK, J. D. (2014). Spatially explicit integrated population models. *Methods Ecol. Evol.* **5** 1351–1360.
- CHANDLER, R. B. and ROYLE, J. A. (2013). Spatially explicit models for inference about density in unmarked or partially marked populations. *Ann. Appl. Stat.* **7** 936–954. [MR3113496](#)
- EDDELBUELTEL, D. and FRANÇOIS, R. (2011). Rcpp: Seamless R and C++ integration. *J. Stat. Softw.* **40** 1–18.
- EDDELBUELTEL, D. and SANDERSON, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Comput. Statist. Data Anal.* **71** 1054–1063.
- FEWSTER, R. M., STEVENSON, B. C. and BORCHERS, D. L. (2016). Trace-contrast models for capture–recapture without capture histories. *Statist. Sci.* **31** 245–258. [MR3506103](#)
- KALLE, R., RAMESH, T., QURESHI, Q. and SANKAR, K. (2011). Density of tiger and leopard in a tropical deciduous forest of Mudumalai Tiger Reserve, southern India, as estimated using photographic capture–recapture sampling. *Acta Theriol.* **56** 335–342.
- KELLY, M. J., BETSCH, J., WULTSCH, C., MESA, B. and MILLS, L. S. (2012). Noninvasive sampling for carnivores. In *Carnivore Ecology and Conservation: A Handbook of Techniques* (L. Boitani and R. A. Powell, eds.) 47–69. Oxford Univ. Press, New York.
- KING, R., MCCLINTOCK, B. T., KIDNEY, D. and BORCHERS, D. (2016). Capture–recapture abundance estimation using a semi-complete data likelihood approach. *Ann. Appl. Stat.* **10** 264–285. [MR3480496](#)
- LINK, W. A., YOSHIKAZI, J., BAILEY, L. L. and POLLOCK, K. H. (2010). Uncovering a latent multinomial: Analysis of mark–recapture data with misidentification. *Biometrics* **66** 178–185. [MR2756704](#)
- MCCLINTOCK, B. T. (2015). Multimark: An R package for analysis of capture–recapture data consisting of multiple “noninvasive” marks. *Ecol. Evol.* **5** 4920–4931.
- MCCLINTOCK, B. T., CONN, P. B., ALONSO, R. S. and CROOKS, K. R. (2013). Integrated modeling of bilateral photo-identification data in mark–recapture analyses. *Ecology* **94** 1464–1471.
- NAIR, T., THORBJARNARSON, J. B., AUST, P. and KRISHNASWAMY, J. (2012). Rigorous gharial population estimation in the Chambal: Implications for conservation and management of a globally threatened crocodilian. *J. Appl. Ecol.* **49** 1046–1054.
- OTIS, D. L., BURNHAM, K. P., WHITE, G. C. and ANDERSON, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs* **62** 3–135.
- REICH, B. J. and GARDNER, B. (2014). A spatial capture–recapture model for territorial species. *Environmetrics* **25** 630–637. [MR3295555](#)
- ROYLE, J. A. (2009). Analysis of capture–recapture models with individual covariates using data augmentation. *Biometrics* **65** 267–274.
- ROYLE, J. A., DORAZIO, R. M. and LINK, W. A. (2007). Analysis of multinomial models with unknown index using data augmentation. *J. Comput. Graph. Statist.* **16** 67–85. [MR2345748](#)
- ROYLE, J. A., FULLER, A. K. and SUTHERLAND, C. (2016). Spatial capture–recapture models allowing Markovian transience or dispersal. *Population Ecology* **58** 53–62.
- ROYLE, J. A., CHANDLER, R. B., SOLLMANN, R. and GARDNER, B. (2013). *Spatial Capture–Recapture*. Academic Press, San Diego.
- SOLLMANN, R., GARDNER, B., PARSONS, A. W., STOCKING, J. J., MCCLINTOCK, B. T., SIMONS, T. R., POLLOCK, K. H. and O’CONNELL, A. F. (2013a). A spatial Mark–resight model augmented with telemetry data. *Ecology* **94** 553–559.
- SOLLMANN, R., TÔRRES, N. M., FURTADO, M. M., DE ALMEIDA JÁCOMO, A. T., PALOMARES, F., ROQUES, S. and SILVEIRA, L. (2013b). Combining camera-trapping and noninvasive genetic data in a spatial capture–recapture framework improves density estimates for the jaguar. *Biological Conservation* **167** 242–247.

- SRIVATHSA, A., PARAMESHWARAN, R., SHARMA, S. and KARANTH, K. U. (2015). Estimating population sizes of leopard cats in the Western Ghats using camera surveys. *J. Mammal.* **96** 742–750.
- SUN, C. C., FULLER, A. K. and ROYLE, J. A. (2014). Trap configuration and spacing influences parameter estimates in spatial capture–recapture models. *PLoS ONE* **9** e88025.
- WANG, S. W. and MACDONALD, D. W. (2009). The use of camera traps for estimating tiger and leopard populations in the high altitude mountains of Bhutan. *Biological Conservation* **142** 606–613.
- WILSON, B., HAMMOND, P. S. and THOMPSON, P. M. (1999). Estimating size and assessing trends in a coastal bottlenose dolphin population. *Ecol. Appl.* **9** 288–300.
- WIMMER, J., TOWSEY, M., ROE, P. and WILLIAMSON, I. (2013). Sampling environmental acoustic recordings to determine bird species richness. *Ecol. Appl.* **23** 1419–1428.
- WRIGHT, J. A., BARKER, R. J., SCHOFIELD, M. R., FRANTZ, A. C., BYROM, A. E. and GLEESON, D. M. (2009). Incorporating genotype uncertainty into mark–recapture-type models for estimating abundance using DNA samples. *Biometrics* **65** 833–840.
- WULTSCH, C., WAITS, L. P. and KELLY, M. J. (2014). Noninvasive individual and species identification of jaguars (*Panthera onca*), pumas (*Puma concolor*) and ocelots (*Leopardus pardalis*) in Belize, Central America using cross-species microsatellites and faecal DNA. *Mol. Ecol. Res.* **14** 1171–1182.

SIMULTANEOUS MODELLING OF MOVEMENT, MEASUREMENT ERROR, AND OBSERVER DEPENDENCE IN MARK-RECAPTURE DISTANCE SAMPLING: AN APPLICATION TO ARCTIC BIRD SURVEYS¹

BY PAUL B. CONN AND RAY T. ALISAUSKAS

NOAA and Environment and Climate Change Canada

Mark-recapture distance sampling is a promising method for surveying bird populations from aircraft in open landscapes. However, commonly available distance sampling estimators require that distances to target animals are made without error and that animals are stationary while sampling is being conducted. Motivated by a recent bird survey where these requirements were routinely violated, we describe a marginal likelihood framework for estimating abundance from double-observer data that can accommodate movement and measurement error when observations are made consecutively (as with front and rear observers), when animals are uniformly distributed during detection by the first observer, and when detections consist of both moving and stationary animals. Assuming that all animals are subject to measurement error and that some animals can move between detections, we integrate over unknown animal locations to construct a marginal likelihood for detection, movement, and measurement error parameters. Estimates of animal abundance are then obtained using a modified Horvitz–Thompson-like estimator. In addition, unmodelled heterogeneity in detection probability can be accommodated through observer dependence parameters. Using simulation, we show that our approach yields low bias compared to approaches that ignore movement and/or measurement error, including in cases where there is considerable detection heterogeneity. Applying our approach to data from a double-observer waterfowl helicopter survey in northern Canada, we are able to estimate bird density accounting for movement and measurement error and corrected for observer heterogeneity. Our approach appears promising for generating unbiased estimates of bird abundance necessary for reliable conservation and management.

REFERENCES

- ALHO, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics* **46** 623–635. [MR1085811](#)
- ALISAUSKAS, R. T. and CONN, P. B. (2018). Evaluating detectability of Arctic waterfowl populations in double-observer helicopter surveys. To be submitted to the Euring 2017 Proceedings.
- BORCHERS, D. L. and EFFORD, M. G. (2008). Spatially explicit maximum likelihood methods for capture-recapture studies. *Biometrics* **64** 377–385. [MR2432407](#)
- BORCHERS, D. L. and MARQUES, T. A. (2017). From distance sampling to spatial capture–recapture. *AStA Adv. Stat. Anal.* **101** 475–494. [MR3712409](#)

Key words and phrases. Aerial survey, double-observer, mark-recapture distance sampling, measurement error, movement, point independence.

- BORCHERS, D. L., ZUCCHINI, W. and FEWSTER, R. M. (1998). Mark-recapture models for line transect surveys. *Biometrics* **54** 1207–1220.
- BORCHERS, D. L., LAAKE, J. L., SOUTHWELL, C. and PAXTON, C. G. M. (2006). Accomodating unmodeled heterogeneity in double-observer distance sampling surveys. *Biometrics* **62** 372–378.
- BORCHERS, D., MARQUES, T., GUNNLAUGSSON, T. and JUPP, P. (2010). Estimating distance sampling detection functions when distances are measured with errors. *J. Agric. Biol. Environ. Stat.* **15** 346–361.
- BORCHERS, D. L., STEVENSON, B. C., KIDNEY, D., THOMAS, L. and MARQUES, T. A. (2015). A unifying model for capture-recapture and distance sampling surveys of wildlife populations. *J. Amer. Statist. Assoc.* **110** 195–204.
- BUCKLAND, S. T., LAAKE, J. L. and BORCHERS, D. L. (2010). Double-observer line transect methods: Levels of independence. *Biometrics* **66** 169–177. [MR2756703](#)
- BUCKLAND, S. T. and TURNOCK, B. J. (1992). A robust line transect method. *Biometrics* 901–909.
- BUCKLAND, S. T., ANDERSON, D. R., BURNHAM, K. P., LAAKE, J. L., BORCHERS, D. L. and THOMAS, L. (2001). *Introduction to Distance Sampling: Estimating the Abundance of Biological Populations*. Oxford Univ. Press, Oxford, United Kingdom.
- BUCKLAND, S. T., ANDERSON, D. R., BURNHAM, K. P., LAAKE, J. L., BORCHERS, D. L. and THOMAS, L. (2004). *Advanced Distance Sampling*. Oxford Univ. Press, Oxford, United Kingdom.
- BURNHAM, K. P., ANDERSON, D. R. and LAAKE, J. L. (1980). Estimation of density for line transect sampling of biological populations. *Wildl. Monogr.* **72** 7–202.
- BURNHAM, K. P., ANDERSON, D. R., WHITE, G. C., BROWNIE, C. and POLLOCK, K. H. (1987). Design and analysis methods for fish survival experiments based on release-recapture. *Am. Fish. Soc. Monogr.* **5** 1–437.
- BURNHAM, K. P., BUCKLAND, S. T., LAAKE, J. L., BORCHERS, D. L., MARQUES, T. A., BISHOP, J. R. B. and THOMAS, L. (2004). Further topics in distance sampling. 307–392. Oxford Univ. Press, Oxford, United Kingdom.
- BURT, M. L., BORCHERS, D. L., JENKINS, K. J. and MARQUES, T. A. (2014). Using mark-recapture distance sampling methods on line transect surveys. *Methods Ecol. Evol.* **5** 1180–1191.
- CONN, P. B., LAAKE, J. L. and JOHNSON, D. S. (2012). A hierarchical modeling framework for multiple observer transect surveys. *PLoS ONE* **7** e42294.
- CONN, P. B., VER HOEF, J. M., MCCLINTOCK, B. T., MORELAND, E. E., LONDON, J. M., CAMERON, M. F., DAHLE, S. P. and BOVENG, P. L. (2014). Estimating multi-species abundance using automated detection systems: Ice-associated seals in the eastern Bering Sea. *Methods Ecol. Evol.* **5** 1280–1293.
- GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. Taylor & Francis, London. [MR2027492](#)
- GLENNIE, R., BUCKLAND, S. T. and THOMAS, L. (2015). The effect of animal movement on line transect estimates of abundance. *PLoS ONE* **10** e0121333.
- HIBY, L. and LOVELL, P. (1998). Using aircraft in tandem formation to estimate abundance of harbour porpoise. *Biometrics* **54** 1280–1289.
- HUGGINS, R. M. (1989). On the statistical analysis of capture-recapture experiments. *Biometrika* **76** 133–140.
- HUGGINS, R. M. (1990). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics* **47** 725–732.
- KONEFF, M. D., ROYLE, J. A., OTTO, M. C., WORTHAM, J. S. and BIDWELL, J. K. (2008). A double-observer method to estimate detection rate during aerial waterfowl surveys **72** 1641–1649.
- LAAKE, J. L. (2013). RMark: An R interface for analysis of capture-recapture data with MARK. AFSC Processed Rep. No. 2013-01, Alaska Fish. Sci. Cent., NOAA, Natl. Mar. Fish. Serv., Seattle, WA.

- LAAKE, J. L. and BORCHERS, D. L. (2004). Methods for incomplete detection at distance zero. In *Advanced Distance Sampling* (S. T. Buckland, D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers and L. Thomas, eds.) 108–189. Oxford Univ. Press, Oxford, United Kingdom.
- MACKENZIE, D. I. and CLEMENT, D. (2016). Accounting for lack of independence and partial overlap of observation zones in line-transect mark-recapture distance sampling. *J. Agric. Biol. Environ. Stat.* **21** 41–57.
- R DEVELOPMENT CORE TEAM (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- ROYLE, J. A., DORAZIO, R. M. and LINK, W. A. (2007). Analysis of multinomial models with unknown index using data augmentation. *J. Comput. Graph. Statist.* **16** 1–19.
- ROYLE, J. A., CHANDLER, R. B., SOLLMANN, R. and GARDNER, B. (2013). *Spatial Capture-Recapture*. Academic Press, San Diego.
- SCHWEDER, T., SKAUG, H. J., LANGAAS, M. and DIMAKOS, X. K. (1999). Simulated likelihood methods for complex double-platform line transect surveys. *Biometrics* **55** 678–687.
- WHITE, G. C. and BURNHAM, K. P. (1999). Program MARK: Survival estimation from populations of marked animals. *Bird Study* **46** Supplement, 120–138.

AN EMPIRICAL STUDY OF THE MAXIMAL AND TOTAL INFORMATION COEFFICIENTS AND LEADING MEASURES OF DEPENDENCE

BY DAVID N. RESHEF^{*,†,§,1}, YAKIR A. RESHEF^{†,‡,1,2},
 PARDIS C. SABETI^{†,**,3} AND MICHAEL MITZENMACHER^{†,**,4}

*Massachusetts Institute of Technology** and *Harvard University*[†]

In exploratory data analysis, we are often interested in identifying promising pairwise associations for further analysis while filtering out weaker ones. This can be accomplished by computing a measure of dependence on all variable pairs and examining the highest-scoring pairs, provided the measure of dependence used assigns similar scores to equally noisy relationships of different types. This property, called *equitability* and previously formalized, can be used to assess measures of dependence along with the power of their corresponding independence tests and their runtime.

Here we present an empirical evaluation of the equitability, power against independence, and runtime of several leading measures of dependence. These include the two recently introduced and simultaneously computable statistics MIC_e , whose goal is equitability, and TIC_e , whose goal is power against independence.

Regarding equitability, our analysis finds that MIC_e is the most equitable method on functional relationships in most of the settings we considered. Regarding power against independence, we find that TIC_e and Heller and Gorfine's S^{DDP} share state-of-the-art performance, with several other methods achieving excellent power as well. Our analyses also show evidence for a trade-off between power against independence and equitability consistent with recent theoretical work. Our results suggest that a fast and useful strategy for achieving a combination of power against independence and equitability is to filter relationships by TIC_e and then to rank the remaining ones using MIC_e . We confirm our findings on a set of data collected by the World Health Organization.

REFERENCES

- ALGEO, T. J. and LYONS, T. W. (2006). Mo–total organic carbon covariation in modern anoxic marine environments: Implications for analysis of paleoredox and paleohydrographic conditions. *Paleoceanography* **21** PA1016.
- BREIMAN, L. and FRIEDMAN, J. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–598.
- BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA. [MR0726392](#)

Key words and phrases. Maximal information coefficient, equitability, total information coefficient, statistical power, measures of dependence.

- CASPI, A., SUGDEN, K., MOFFITT, T. E., TAYLOR, A., CRAIG, I. W., HARRINGTON, H., MCCLAY, J., MILL, J., MARTIN, J., BRAITHWAITE, A. and POULTON, R. (2003). Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science* **301** 386–389.
- CLAYTON, R. N. and MAYEDA, T. K. (1996). Oxygen isotope studies of achondrites. *Geochim. Cosmochim. Acta* **60** 1999–2017.
- DING, A. A. and LI, Y. (2013). Copula correlation: An equitable dependence measure and extension of pearson’s correlation. Preprint. Available at [arXiv:1312.7214](https://arxiv.org/abs/1312.7214).
- EMILSSON, V., THORLEIFSSON, G., ZHANG, B., LEONARDSON, A. S., ZINK, F., ZHU, J., CARLSON, S., HELGASON, A., BRAGI WALTERS, G., GUNNARSDOTTIR, S. et al. (2008). Genetics of gene expression and its effect on disease. *Nature* **452** 423–428.
- GILL, T. ET AL. (2002). Obesity in the pacific: Too big to ignore. World Health Organization Regional Office for the Western Pacific, Secretariat of the Pacific Community.
- GORFINE, M., HELLER, R. and HELLER, Y. (2012). Comment on “Detecting novel associations in large data sets.” Unpublished. Available at <http://www.math.tau.ac.il/~ruheller/Papers/science6.pdf>.
- GRETTON, A., BOUSQUET, O., SMOLA, A. and SCHÖLKOPF, B. (2005). Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory* 63–77. Springer, Berlin.
- GRETTON, A., FUKUMIZU, K., TEO, C. H., LE, S., SCHÖLKOPF, B. and SMOLA, A. J. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems* 585–592.
- HELLER, R., HELLER, Y. and GORFINE, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100** 503–510. [MR3068450](https://doi.org/10.1093/biomet/asx045)
- HELLER, R., HELLER, Y., KAUFMAN, S., BRILL, B. and GORFINE, M. (2016). Consistent distribution-free k -sample and independence tests for univariate random variables. *J. Mach. Learn. Res.* **17** 1–54.
- HOEFFDING, W. (1948). A non-parametric test of independence. *Ann. Math. Stat.* 546–557.
- HUO, X. and SZEKELY, G. J. (2014). Fast computing for distance covariance. Preprint. Available at [arXiv:1410.1503](https://arxiv.org/abs/1410.1503).
- JAKKOLA, T. S. and HAUSSLER, D. (1999). Probabilistic kernel regression models. In *AISTATS*.
- JIANG, B., YE, C. and LIU, J. S. (2015). Nonparametric k -sample tests via dynamic slicing. *J. Amer. Statist. Assoc.* **110** 642–653.
- KINNEY, J. B. and ATWAL, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci. USA* **111** 3354–3359. [MR3200177](https://doi.org/10.1073/pnas.1312777111)
- KRASKOV, A., STOGBAUER, H. and GRASSBERGER, P. (2004). Estimating mutual information. *Phys. Rev. E* **69** 066138.
- LINFOOT, E. H. (1957). An informational measure of correlation. *Inf. Control* **1** 85–89.
- LOPEZ-PAZ, D., HENNIG, P. and SCHÖLKOPF, B. (2013). The randomized dependence coefficient. In *Advances in Neural Information Processing Systems* 1–9.
- MOON, Y.-I., RAJAGOPALAN, B. and LALL, U. (1995). Estimation of mutual information using kernel density estimators. *Phys. Rev. E* **52** 2318–2321.
- MURRELL, B., MURRELL, D. and MURRELL, H. (2014). R2-equitability is satisfiable. *Proc. Natl. Acad. Sci. USA* **111** E2160–E2160. Available at <http://www.pnas.org/content/early/2014/04/29/1403623111>.
- PANINSKI, L. (2003). Estimation of entropy and mutual information. *Neural Comput.* **15** 1191–1253.
- RÉNYI, A. (1959). On measures of dependence. *Acta Math. Hungar.* **10** 441–451.
- RESHEF, D. N., RESHEF, Y. A., SABETI, P. C. and MITZENMACHER, M. (2018a). Appendix to “An empirical study of the maximal and total information coefficients and leading measures of dependence.” DOI:[10.1214/17-AOAS1093SUPPA](https://doi.org/10.1214/17-AOAS1093SUPPA).
- RESHEF, D. N., RESHEF, Y. A., SABETI, P. C. and MITZENMACHER, M. (2018b). Supplement to “An empirical study of the maximal and total information coefficients and leading measures of dependence.” DOI:[10.1214/17-AOAS1093SUPPB](https://doi.org/10.1214/17-AOAS1093SUPPB).

- RESHEF, D. N., RESHEF, Y. A., FINUCANE, H. K., GROSSMAN, S. R., MCVEAN, G., TURN-BAUGH, P. J., LANDER, E. S., MITZENMACHER, M. and SABETI, P. C. (2011). Detecting novel associations in large data sets. *Science* **334** 1518–1524.
- RESHEF, D., RESHEF, Y., MITZENMACHER, M. and SABETI, P. (2013). Equitability analysis of the maximal information coefficient, with comparisons. Preprint. Available at [arXiv:1301.6314](https://arxiv.org/abs/1301.6314).
- RESHEF, D. N., RESHEF, Y. A., MITZENMACHER, M. and SABETI, P. C. (2014). Cleaning up the record on the maximal information coefficient and equitability. *Proc. Natl. Acad. Sci. USA* **111** E3362–E3363. Available at <http://www.pnas.org/content/early/2014/08/07/1408920111>.
- RESHEF, Y. A., RESHEF, D. N., SABETI, P. C. and MITZENMACHER, M. (2015). Equitability, interval estimation, and statistical power. Available at [arXiv:1505.02212](https://arxiv.org/abs/1505.02212).
- RESHEF, Y. A., RESHEF, D. N., FINUCANE, H. K., SABETI, P. C. and MITZENMACHER, M. (2016). Measuring dependence powerfully and equitably. *J. Mach. Learn. Res.* **17** Paper No. 212, 63. [MR3595146](#)
- SEJDINOVIC, D., SRIPERUMBUDUR, B., GRETTON, A. and FUKUMIZU, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.* **41** 2263–2291. [MR3127866](#)
- SIMON, N. and TIBSHIRANI, R. (2012). Comment on “Detecting novel associations in large data sets”. Unpublished. Available at <http://statweb.stanford.edu/tibs/reshef/comment.pdf>.
- SPEED, T. (2011). A correlation for the 21st century. *Science* **334** 1502–1503.
- SZEKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.
- WANG, X., JIANG, B. and LIU, J. S. (2017). Generalized R-squared for detecting dependence. *Biometrika* **104** 129–139. [MR3626486](#)

COVARIATE BALANCING PROPENSITY SCORE FOR A CONTINUOUS TREATMENT: APPLICATION TO THE EFFICACY OF POLITICAL ADVERTISEMENTS

BY CHRISTIAN FONG, CHAD HAZLETT, AND KOSUKE IMAI

*Stanford University, University of California, Los Angeles and
 Princeton University*

Propensity score matching and weighting are popular methods when estimating causal effects in observational studies. Beyond the assumption of unconfoundedness, however, these methods also require the model for the propensity score to be correctly specified. The recently proposed covariate balancing propensity score (CBPS) methodology increases the robustness to model misspecification by directly optimizing sample covariate balance between the treatment and control groups. In this paper, we extend the CBPS to a continuous treatment. We propose the covariate balancing generalized propensity score (CBGPS) methodology, which minimizes the association between covariates and the treatment. We develop both parametric and non-parametric approaches and show their superior performance over the standard maximum likelihood estimation in a simulation study. The CBGPS methodology is applied to an observational study, whose goal is to estimate the causal effects of political advertisements on campaign contributions. We also provide open-source software that implements the proposed methods.

REFERENCES

- ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267. [MR2194325](#)
- BOYD, C. L., EPSTEIN, L. and MARTIN, A. D. (2010). Untangling the causal effects of sex on judging. *Amer. J. Polit. Sci.* **54** 389–411.
- CHAN, K. C. G., YAM, S. C. P. and ZHANG, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 673–700. [MR3506798](#)
- DE, P. K. and RATHA, D. (2012). Impact of remittances on household income, asset, and human capital: Evidence from Sri Lanka. *Migr. Dev.* **1** 163–179.
- DONOHUE III, J. J. and HO, D. E. (2007). The impact of damage caps on malpractice claims: Randomization inferences with difference-in-differences. *J. Empir. Leg. Stud.* **4** 69–102.
- FAN, J., IMAI, K., LIU, H., NING, Y. and YANG, X. (2016). Improving covariate balancing propensity score: A doubly robust and efficient approach. Technical report, Princeton Univ.
- FONG, C., HAZLETT, C. and IMAI, K. (2018). Replication data for: Covariate balancing propensity score for a continuous treatment. DOI:[10.7910/DVN/AIF4PI](#).
- FONG, C., RATKOVIC, M., HAZLETT, C. and IMAI, K. (2017). CBPS: R package for covariate balancing propensity score. Available at the Comprehensive R Archive Network (CRAN): <https://CRAN.R-project.org/package=CBPS>.

Key words and phrases. Causal inference, covariate balance, generalized propensity score, inverse-probability weighting, treatment effect.

- GRAHAM, B. S., PINTO, C. and EGEL, D. (2012). Inverse probability tilting for moment condition models with missing data. *Rev. Econ. Stud.* **79** 1053–1079.
- HAINMUELLER, J. (2012). Entropy balancing for causal effects: Multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20** 25–46.
- HARDER, V. S., STUART, E. A. and ANTHONY, J. C. (2008). Adolescent cannabis problems and young adult depression: Male–female stratified propensity score analyses. *Am. J. Epidemiol.* **168** 592–601.
- HAZLETT, C. (2016). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. Technical report, Depts. Statistics and Political Science, Univ. California Los Angeles.
- HIRANO, K. and IMBENS, G. W. (2004). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family* 73–84. Wiley, New York.
- HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. [MR1995826](#)
- 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.
- IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 243–263. [MR3153941](#)
- IMAI, K. and RATKOVIC, M. (2015). Robust estimation of inverse probability weights for marginal structural models. *J. Amer. Statist. Assoc.* **110** 1013–1023.
- IMAI, K. and VAN DYK, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* **99** 854–866.
- IMBENS, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710.
- IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.
- KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. [MR2420458](#)
- MCCAFFREY, D. F., RIDGEWAY, G. and MORRAL, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* **9** 403–425.
- NEWBY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics, Vol. IV. Handbooks in Econom.* **2** 2111–2245. North-Holland, Amsterdam. [MR1315971](#)
- NIELSEN, R. A., FINDLEY, M. G., DAVIS, Z. S., CANDLAND, T. and NIELSON, D. L. (2011). Foreign aid shocks as a cause of violent armed conflict? *Amer. J. Polit. Sci.* **55** 219–232.
- OWEN, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton, FL.
- ROBINS, J. M., HERNÁN, M. Á. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.
- ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.
- ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.
- ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.
- RUBIN, D. B. (1990). Comments on “On the application of probability theory to agricultural experiments. Essay on principles. Section 9” by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statist. Sci.* **5** 472–480.

- SMITH, J. A. and TODD, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *J. Econometrics* **125** 305–353.
- STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. [MR2741812](#)
- TAN, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97** 661–682. [MR2672490](#)
- URBAN, C. and NIEBLER, S. (2014). Dollars on the sidewalk: Should US presidential candidates advertise in uncontested states? *Amer. J. Polit. Sci.* **58** 322–336.
- ZHAO, Q. (2016). Covariate balancing propensity score by tailored loss functions. Technical report. Dept. Statistics, Stanford Univ.
- ZHAO, Q. and PERCIVAL, D. (2017). Entropy balancing is doubly robust. *J. Causal Inference* **5** 20160010.
- ZHU, Y., COFFMAN, D. L. and GHOSH, D. (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. *J. Causal Inference* **3** 25–40.
- ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* **110** 910–922.

INTEGRATIVE EXPLORATION OF LARGE HIGH-DIMENSIONAL DATASETS¹

BY CHRISTOPHER PARDY^{*,†}, SALLY GALBRAITH^{*} AND SUSAN R. WILSON[†]

University of New South Wales^{} and Australian National University[†]*

Large, high-dimensional datasets containing different types of variables are becoming increasingly common. For exploring such data, there is a need for integrated methods. For example, a single genomic experiment can contain large quantities of different types of data (including clinical data) that make it a challenge to coherently describe the patterns of variability within and between the inter-related datasets. Mutual information (MI) is a widely used information theoretic dependency measure that also can identify non-linear and nonmonotonic associations. First, we develop a computationally efficient implementation of MI between a discrete and a continuous variable. This implementation allows us to apply a coherent approach to all comparisons arising from continuous and categorical data. As commonly applied, MI can have high levels of bias. So we present a novel development of mutual information (MI) that reduces the bias, and that we term bias corrected mutual information (BCMI). Further, BCMI is useful as an association measure that can be incorporated in subsequent analyses such as clustering and visualisation procedures.

To demonstrate our approach, a genomic dataset is re-examined. This dataset contains single nucleotide polymorphisms (SNPs, a discrete variable), gene expression levels and clinical data (all continuous variables). Our approach allows us to integrate these different types of data by exploring associations both within and between these types of variables.

REFERENCES

- CHU, J., WEISS, S. T., CAREY, V. J. and RABY, B. A. (2009). A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. *BMC Syst. Biol.* **3** 55.
- COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley-Interscience, Hoboken, NJ. [MR2239987](#)
- DAWY, Z., GOEBEL, B., HAGENAUER, J., ANDREOLI, C., MEITINGER, T. and MUELLER, J. C. (2006). Gene mapping and marker clustering using Shannon’s mutual information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **3** 47–56.
- DUTKOWSKI, J., KRAMER, M., SURMA, M. A., BALAKRISHNAN, R., CHERRY, J. M., KROGAN, N. J. and IDEKER, T. (2013). A gene ontology inferred from molecular networks. *Nat. Biotechnol.* **31** 38–45.
- EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics **38**. SIAM, Philadelphia, PA. [MR0659849](#)

Key words and phrases. Mutual information, data integration, exploration, mixed-types of variables, continuous, categorical.

- EFRON, B. (2010). *Large-Scale Inference. Empirical Bayes Methods For Estimation, Testing, and Prediction. Institute of Mathematical Statistics (IMS) Monographs 1*. Cambridge Univ. Press, Cambridge. [MR2724758](#)
- EFRON, B. and GONG, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *Amer. Statist.* **37** 36–48. [MR0694281](#)
- FULLER, T. F., GHAZALPOUR, A., ATEN, J. E., DRAKE, T. A., LUSIS, A. J. and HORVATH, S. (2007). Weighted gene coexpression network analysis strategies applied to mouse weight. *Mamm. Genome* **18** 463–472.
- GHAZALPOUR, A., DOSS, S., ZHANG, B., WANG, S., PLAISIER, C., CASTELLANOS, R., BROZELL, A., SCHAT, E. E., DRAKE, T. A., LUSIS, A. J. et al. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* **2** e130.
- HALL, P. and MILLER, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Statist.* **18** 533–550. [MR2751640](#)
- HALL, P. and MILLER, H. (2011). Determining and depicting relationships among components in high-dimensional variable selection. *J. Comput. Graph. Statist.* **20** 988–1006. [MR2878959](#)
- HINTZE, J. L. and NELSON, R. D. (1998). Violin plots: A box plot-density trace synergism. *Amer. Statist.* **52** 181–184.
- KRASKOV, A., STÖGBAUER, H. and GRASSBERGER, P. (2004). Estimating mutual information. *Phys. Rev. E* (3) **69** 066138. [MR2096503](#)
- LAIRD, N. M. and LANGE, C. (2011). *The Fundamentals of Modern Statistical Genetics*. Springer, New York. [MR2762582](#)
- MARRON, J. S. and PADGETT, W. J. (1987). Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *Ann. Statist.* **15** 1520–1535. [MR0913571](#)
- PADGETT, W. J. and MCNICHOLS, D. T. (1984). Nonparametric density estimation from censored data. *Comm. Statist. Theory Methods* **13** 1581–1611. [MR0752184](#)
- PARDY, C., GALBRAITH, S. and WILSON, S. R. (2018). Supplement to “Integrative exploration of large high-dimensional datasets.” DOI:[10.1214/17-AOAS1055SUPP](#).
- PRINCIPE, J. C. (2010). *Information Theoretic Learning. Renyi's Entropy and Kernel Perspectives. Information Science and Statistics*. Springer, New York. [MR2598963](#)
- QIU, P., GENTLES, A. J. and PLEVritis, S. K. (2009). Fast calculation of pairwise mutual information for gene regulatory network reconstruction. *Comput. Methods Programs Biomed.* **94** 177–180.
- QUENOUILLE, M. H. (1956). Notes on bias in estimation. *Biometrika* **43** 353–360. [MR0081040](#)
- RESHEF, D. N., RESHEF, Y. A., FINUCANE, H. K., GROSSMAN, S. R., McVEAN, G., TURNBAUGH, P. J., LANDER, E. S., MITZENMACHER, M. and SABETI, P. C. (2011). Detecting novel associations in large data sets. *Science* **334** 1518–1524.
- SCHELLHASE, C. and KAUFMANN, G. (2012). Density estimation and comparison with a penalized mixture approach. *Comput. Statist.* **27** 757–777. [MR3041856](#)
- SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. and IDEKER, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13** 2498–2504.
- SHEATHER, S. J. and JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **53** 683–690. [MR1125725](#)
- STEUER, R., KURTHS, J., DAUB, C. O., WEISE, J. and SELBIG, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* **18**(Suppl. 2) S231–S240.
- SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265. [MR2752127](#)
- WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing. Monographs on Statistics and Applied Probability 60*. Chapman and Hall, London. [MR1319818](#)

WANG, S., YEHYA, N., SCHADT, E. E., WANG, H., DRAKE, T. A. and LUSIS, A. J. (2006). Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genet.* **2** e15. DOI:[10.1371/journal.pgen.0020015](https://doi.org/10.1371/journal.pgen.0020015).

ADJUSTMENT OF NONCONFOUNDING COVARIATES IN CASE-CONTROL GENETIC ASSOCIATION STUDIES

BY HONG ZHANG^{*,†,1,2}, NILANJAN CHATTERJEE[‡],
DANIEL RADER[†] AND JINBO CHEN^{†,1}

Fudan University^{}, University of Pennsylvania[†] and
National Institutes of Health[‡]*

It has recently been reported that adjustment of nonconfounding covariates in case-control genetic association analyses may lead to decreased power when the phenotype is rare. This observation contrasts a well-known result for clinical trials where adjustment of baseline variables always leads to increased power for testing randomized treatment effects. In this paper, we propose a unified solution that guarantees increased power through covariate adjustment regardless of whether the phenotype is rare or common. Our method exploits external phenotype prevalence data through a profile likelihood function, and can be applied to fit any commonly used penetrance models including the logistic and probit regression models. Through extensive simulation studies, we showed empirically that the power of our method was indeed higher than available analysis strategies with or without covariate adjustment, and can be considerably higher when the phenotype was common and the covariate effect was strong. We applied the proposed method to analyze a case-control genetic association study on human high density lipoprotein cholesterol level.

REFERENCES

- BENJAMIN, M. N., MANUEL, A. R., BENJAMIN, F. V., DAVID, A., BERNIE, D., MARJU, O., SEKAR, K., SHAUN, M. P., KATHRYN, R. and MARK, J. D. (2011). Testing for an unusual distribution of rare variants. *PLoS Genet.* **7** e1001322.
- CHATTERJEE, N. and CARROLL, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92** 399–418. MR2201367
- EDMONDSON, A. C., BRAUND, P. S., STYLIANOU, I. M., KHERA, A. V., NELSON, C. P., WOLFE, M. L., DEROHANNESSIAN, S. L., KEATING, B. J., QU, L., HE, J., TOBIN, M. D., TOMASZEWSKI, M., BAUMERT, J., KLOPP, N., DÖRING, A., THORAND, B., LI, M., REILLY, M. P., KOENIG, W., SAMANI, N. J. and RADER, D. J. (2011). Dense genotyping of candidate gene loci identifies variants associated with high-density lipoprotein cholesterol. *Circ. Cardiovasc. Genet.* **4** 145–155.
- KUO, C.-L. and FEINGOLD, E. (2010). What's the best statistic for a simple test of genetic association in a case-control study? *Genet. Epidemiol.* **34** 246–253.
- LEE, S., EMOND, M. J., BAMSHAD, M. J., BARNES, K. C., RIEDER, M. J., NICKERSON, D. A., NHLBI GO EXOME SEQUENCING PROJECT—ESP LUNG PROJECT TEAM,

Key words and phrases. Case-control studies, Hardy–Weinberg equilibrium, nonconfounding covariates, profile likelihood.

- CHRISTIANI, D. C., WURFEL, M. M. and LIN, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91** 224–237.
- LI, B. and LEAL, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am. J. Hum. Genet.* **83** 311–321.
- NEUHAUS, J. M. (1998). Estimation efficiency with omitted covariates in generalized linear models. *J. Amer. Statist. Assoc.* **93** 1124–1129. [MR1649206](#)
- NEUHAUS, J. M. and JEWELL, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika* **80** 807–815. [MR1282789](#)
- OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. [MR0946049](#)
- OWEN, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18** 90–120. [MR1041387](#)
- PIRINEN, M., DONNELLY, P. and SPENCER, C. C. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat. Genet.* **44** 848–851.
- QIN, J. and LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22** 300–325. [MR1272085](#)
- QIN, J., ZHANG, H., LI, P., ALBANES, D. and YU, K. (2015). Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika* **102** 169–180. [MR3335103](#)
- STRINGER, S., WRAY, N. R., KAHN, R. S. and DERKS, E. M. (2011). Underestimated effect sizes in GWAS: Fundamental limitations of single SNP analysis for dichotomous phenotypes. *PLoS ONE* **6** e27964.
- SUN, J., ZHENG, Y. and HSU, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genet. Epidemiol.* **37** 334–344.
- WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. et al. (2001). Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *Am. J. Hum. Genet.* **89** 82–93.
- ZAITLEN, N., PAŞANIUC, B., PATTERSON, N., POLLACK, S., VOIGHT, B., GROOP, L., ALTSHULER, D., HENDERSON, B. E., KOLONEL, L. N., LE MARCHAND, L., WATERS, K., HAIMAN, C. A., STRANGER, B. E., DERMITZAKIS, E. T., KRAFT, P. and PRICE, A. L. (2012a). Analysis of case-control association studies with known risk variants. *Bioinformatics* **28** 1729–1737.
- ZAITLEN, N., LINDSTRÖM, S., PASANIUC, B., CORNELIS, M., GENOVESE, G., POLLACK, S., BARTON, A., BICKEBÖLLER, H., BOWDEN, D. W., EYRE, S., FREEDMAN, B. I., FRIEDMAN, D. J., FIELD, J. K., GROOP, L., HAUGEN, A., HEINRICH, J., HENDERSON, B. E., HICKS, P. J., HOCKING, L. J., KOLONEL, L. N., LANDI, M. T., LANGEFELD, C. D., LE MARCHAND, L., MEISTER, M., MORGAN, A. W., RAJI, O. Y., RISCH, A., ROSENBERGER, A., SCHERF, D., STEER, S., WALSHAW, M., WATERS, K. M., WILSON, A. G., WORDSWORTH, P., ZIENOLDDINY, S., TCHETGEN, E. T., HAIMAN, C., HUNTER, D. J., PLENGE, R. M., WORTHINGTON, J., CHRISTIANI, D. C., SCHAUMBERG, D. A., CHASMAN, D. I., ALTSHULER, D., VOIGHT, B., KRAFT, P., PATTERSON, N. and PRICE, A. L. (2012b). Informed conditioning on clinical covariates increases power in case-control association studies. *PLoS Genet.* **8** e1003032.
- ZHANG, M., TSIATIS, A. A. and DAVIDIAN, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* **64** 707–715. [MR2526620](#)
- ZHANG, H., CHATTERJEE, N., RADER, D. and CHEN, J. (2018). Supplement to “Adjustment of nonconfounding covariates in case-control genetic association studies.” DOI:10.1214/17-AOAS1065SUPP.

HOW GAUSSIAN MIXTURE MODELS MIGHT MISS DETECTING FACTORS THAT IMPACT GROWTH PATTERNS¹

BY BRIANNA C. HEGGESETH AND NICHOLAS P. JEWELL

Williams College and University of California, Berkeley

Longitudinal studies play a prominent role in biological, social, and behavioral sciences. Repeated measurements over time facilitate the study of an outcome level, how individuals change over time, and the factors that may impact either or both. A standard approach to modeling childhood growth over time is to use multilevel or mixed effects models to study factors that might play a role in the level and growth over time. However, there has been increased interest in using mixture models, which have inherent grouping structure to more flexibly explain heterogeneity in the longitudinal outcomes, to study growth patterns. While several possible model specifications can be used, these methods generally fail to explicitly group individuals by the shape of their growth pattern separate from level, and thus fail to shed light on the relationships between growth pattern and potential explanatory factors. We illustrate the weaknesses of these methods as they are currently being used. We also propose a pre-processing step that removes the outcome level to focus explicitly on shape, discuss its impact on estimation, and demonstrate its usefulness through a simulation study and with real longitudinal data.

REFERENCES

- AITKIN, M., ANDERSON, D. and HINDE, J. (1981). Statistical modelling of data on teaching styles. *J. Roy. Statist. Soc. Ser. A* **144** 419–461.
- ASPAROUHOV, T. and MUTHÉN, B. (2016). Structural equation models and mixture models with continuous nonnormal skewed distributions. *Struct. Equ. Model.* **23** 1–19.
- BRILLINGER, D. R. (1975). *Time Series: Data Analysis and Theory*. Holt, Rinehart and Winston, New York.
- CARTER, M. A., DUBOIS, L., TREMBLAY, M. S., TALJAARD, M. and JONES, B. L. (2012). Trajectories of childhood weight gain: The relative importance of local environment versus individual social and early life factors. *PLoS ONE* **7** e47065.
- CELEUX, G. and SOROMENHO, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *J. Classification* **13** 195–212.
- CUPUL-UICAB, L. A., HERNÁNDEZ-AVILA, M., TERRAZAS-MEDINA, E. A., PENNELL, M. L. and LONGNECKER, M. P. (2010). Prenatal exposure to the major DDT metabolite 1,1-dichloro-2,2-bis(p-chlorophenyl)ethylene (DDE) and growth in boys from Mexico. *Environ. Res.* **110** 595–603.
- CUPUL-UICAB, L. A., KLEBANOFF, M. A., BROCK, J. W. and LONGNECKER, M. P. (2013). Prenatal exposure to persistent organochlorines and childhood obesity in the US collaborative perinatal project. *Environmental Health Perspectives* **121** 1103–1109.

Key words and phrases. Finite mixture model, longitudinal data analysis, latent variables, growth curves.

- CURRY, H. B. and SCHOENBERG, I. J. (1966). On Pólya frequency functions IV: The fundamental spline functions and their limits. *Journal d'Analyse Mathématique* **17** 71–107.
- D'URSO, P. (2000). Dissimilarity measures for time trajectories. *Stat. Methods Appl.* **9** 53–83.
- DAVIES, C. E., GLONEK, G. F. V. and GILES, L. C. (2015). The impact of covariance misspecification in group-based trajectory models for longitudinal data with non-stationary covariance structure. *Stat. Methods Med. Res.* Preprint. Available online doi:10.1177/0962280215598806.
- DEEKS, S. G., HECHT, F. M., SWANSON, M., ELBEIK, T., LOFTUS, R., COHEN, P. T. and GRANT, R. M. (1999). HIV RNA and CD4 cell count response to protease inhibitor therapy in an urban AIDS clinic: Response to both initial and salvage therapy. *AIDS* **13** 35–43.
- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39** 1–38.
- DE BOOR, C. (1976). Splines as linear combinations of B-splines. A survey. In *Approximation Theory II* (G. G. Lorentz, C. K. Chui and L. L. Schumaker, eds.) 1–47. Academic Press, New York.
- DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer, New York.
- DIALLO, T. M. O., MORIN, A. J. S. and LU, H. (2016). Impact of misspecifications of the latent variance? Covariance and residual matrices on the class enumeration accuracy of growth mixture models. *Struct. Equ. Model.* **23** 507–531.
- DIGGLE, P., HEAGERTY, P., LIANG, K. Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. Oxford Univ. Press, New York.
- EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.
- EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics **38**. SIAM, Philadelphia, PA.
- EROSHEVA, E. A., MATSUEDA, R. L. and TELESKA, D. (2014). Breaking bad: Two decades of life-course data analysis in criminology, developmental psychology, and beyond. *Annual Review of Statistics and Its Application* **1** 301–332.
- ESKENAZI, B., BRADMAN, A., GLADSTONE, E. A., JARAMILLO, S., BIRCH, K. and HOLLAND, N. (2003). CHAMACOS, a longitudinal birth cohort study: Lessons from the fields. *Journal of Children's Health* **1** 3–27.
- ESKENAZI, B., HARLEY, K., BRADMAN, A., WELTZIEN, E., JEWELL, N. P., BARR, D. B., FURLONG, C. E. and HOLLAND, N. T. (2004). Association of in utero organophosphate pesticide exposure and fetal growth and length of gestation in an agricultural population. *Environmental Health Perspectives* **112** 1116–1124.
- ESKENAZI, B., GLADSTONE, E. A., BERKOWITZ, G. S., DREW, C. H., FAUSTMAN, E. M., HOLLAND, N. T., LANPHEAR, B., MEISEL, S. J., PERERA, F. P., RAUH, V. A., SWEENEY, A., WHYATT, R. M. and YOLTON, K. (2005). Methodologic and logistic issues in conducting longitudinal birth cohort studies: Lessons learned from the Centers for Children's Environmental Health and Disease Prevention Research. *Environmental Health Perspectives* **113** 1419–1429.
- EUBANK, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. Dekker, New York, NY.
- EVERITT, B. S. and HAND, D. J. (1981). *Finite Mixture Distributions*. Chapman & Hall, London.
- EVERITT, B. S., LANDAU, S., LEESE, M. and STAHL, D. (2011). *Cluster Analysis*, 5th ed. Wiley, London.
- FENG, Z. D. and MCCULLOCH, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 609–617.
- FRALEY, C. and RAFTERY, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J.* **41** 578–588.
- GARDEN, F. L., MARKS, G. B., SIMPSON, J. M. and WEBB, K. L. (2012). Body mass index (BMI) trajectories from birth to 11.5 years: Relation to early life food intake. *Nutrients* **4** 1382–1398.
- GRAY, G. (1994). Bias in misspecified mixtures. *Biometrics* **50** 457–470.
- GRUN, B. and LEISCH, F. (2008). FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *J. Stat. Softw.* **28** 1–35.

- HEGGESETH, B. (2018a). Supplement to “How Gaussian mixture models might miss detecting factors that impact growth patterns.” DOI:[10.1214/17-AOAS1066SUPPA](https://doi.org/10.1214/17-AOAS1066SUPPA).
- HEGGESETH, B. (2018b). Supplement to “How Gaussian mixture models might miss detecting factors that impact growth patterns.” DOI:[10.1214/17-AOAS1066SUPPB](https://doi.org/10.1214/17-AOAS1066SUPPB).
- HEGGESETH, B. C. and JEWELL, N. P. (2013). The impact of covariance misspecification in multivariate Gaussian mixtures on estimation and inference: An application to longitudinal modeling. *Stat. Med.* **32** 2790–2803.
- HEO, M., FAITH, M. S., MOTT, J. W., GORMAN, B. S., REDDEN, D. T. and ALLISON, D. B. (2003). Hierarchical linear models for the development of growth curves: An example with body mass index in overweight/obese adults. *Stat. Med.* **22** 1911–1942.
- HUANG, Y., CHEN, J. and YIN, P. (2017). Hierarchical mixture models for longitudinal immunologic data with heterogeneity, non-normality, and missingness. *Stat. Methods Med. Res.* **26** 223–247.
- HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.
- JENNRICH, R. I. and SCHLUCHTER, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42** 805–820.
- JONES, B. L., NAGIN, D. S. and ROEDER, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociol. Methods Res.* **29** 374–393.
- LEISCH, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *J. Stat. Softw.* **11** 1–18.
- LU, X. and HUANG, Y. (2014). Bayesian analysis of nonlinear mixed-effects mixture models for longitudinal data with heterogeneity and skewness. *Stat. Med.* **33** 2830–2849.
- MCLACHLAN, G. J. and BASFORD, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Dekker, New York.
- MCLACHLAN, G. J. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York.
- MENDEZ, M. A., GARCIA-ESTEBAN, R., GUXENS, M., VRIJHEID, M., KOGEVINAS, M., GOÑI, F., FOCHS, S. and SUNYER, J. (2011). Prenatal organochlorine compound exposure, rapid weight gain, and overweight in infancy. *Environmental Health Perspectives* **119** 272–278.
- MÖLLER-LEVET, C., KLAWONN, F., CHO, K. H. and WOLKENHAUER, O. (2003). Fuzzy clustering of short time-series and unevenly distributed sampling points. In *Proceedings of the Fifth International Conference on Intelligent Data Analysis* (M. R. Berthold, H.-J. Lenz, E. Bradley and C. Borgelt, eds.) 330–340.
- MORIN, A. J. S. and MARSH, H. W. (2015). Disentangling shape from level effects in person-centered analyses: An illustration based on university teachers? Multidimensional profiles of effectiveness. *Struct. Equ. Model.* **22** 39–59.
- MORIN, A. J. S., MAÍANO, C., MARSH, H. W., NAGENGAST, B. and JANOSZ, M. (2013). School life and adolescents’ self-esteem trajectories. *Child Dev.* **84** 1967–1988.
- MUTHÉN, B. and ASPAROUHOV, T. (2009). Multilevel regression mixture analysis. *J. Roy. Statist. Soc. Ser. A* **172** 639–657.
- MUTHÉN, L. K. and MUTHÉN, B. O. (1998–2010). *Mplus User’s Guide*, 6th ed., Los Angeles.
- MUTHÉN, B. and SHEDDEN, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55** 463–469.
- MUTHÉN, B., BROWN, C. H., MASYN, K., JO, B., KHOO, S. T., YANG, C. C., WANG, C. P., KELLAM, S. G., CARLIN, J. B. and LIAO, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics* **3** 459–475.
- NAGIN, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods* **4** 139–157.
- NAGIN, D. S. and ODGERS, C. L. (2010a). Group-based trajectory modeling (nearly) two decades later. *J. Quant. Criminol.* **26** 445–453.
- NAGIN, D. S. and ODGERS, C. L. (2010b). Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology* **6** 109–138.

- PARK, T., YI, S.-G., KANG, S.-H., LEE, S., LEE, Y.-S. and SIMON, R. (2003). Evaluation of normalization methods for microarray data. *BMC Bioinform.* **4** 1–13.
- PEARSON, K. (1894). Contributions to the mathematical theory of evolution. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **185** 71–110.
- PICKLES, A. and CROUDACE, T. (2010). Latent mixture models for multivariate and longitudinal outcomes. *Stat. Methods Med. Res.* **19** 271–289.
- PROUST-LIMA, C., PHILIPPS, V., DIAKITE, A. and LIQUET, B. (2014). lcm: Estimation of extended mixed models using latent classes and latent processes. R package version 1.6.4.
- PRYOR, L. E., TREMBLAY, R. E., BOIVIN, M., TOUCHETTE, E., DUBOIS, L., GENOLINI, L., XUECHENG, C., FALISSARD, B. and CÔTÉ, S. M. (2011). Developmental trajectories of body mass index in early childhood and their risk factors: An 8-year longitudinal study. *Archives of Pediatrics & Adolescent Medicine* **165** 906–912.
- RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.
- RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.* **11** 735–757.
- SCHLATTMANN, P. and BÖHNING, D. (1997). On Bayesian analysis of mixtures with an unknown number of components. Contribution to a paper by S. Richardson and PJ Green. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **59** 782–783.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.
- SHUMWAY, R. H. and STOFFER, D. S. (2010). *Time Series Analysis and Its Applications: With R Examples*. Springer Science & Business Media, New York.
- SINGER, J. D. and WILLETT, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford Univ. Press, New York, NY.
- TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.
- VALVI, D., MENDEZ, M. A., MARTINEZ, D., GRIMALT, J. O., TORRENT, M., SUNYER, J. and VRIJHEID, M. (2012). Prenatal concentrations of polychlorinated biphenyls, DDE, and DDT and overweight in children: A prospective birth cohort study. *Environmental Health Perspectives* **120** 451–457.
- WARNER, M., AGUILAR SCHALL, R., HARLEY, K. G., BRADMAN, A., BARR, D. and ESKENAZI, B. (2013). In utero DDT and DDE exposure and obesity status of 7-year-old Mexican-American children in the CHAMACOS cohort. *Environmental Health Perspectives* **121** 631–636.
- WARNER, M., WESSELINK, A., HARLEY, K. G., BRADMAN, A., KOGUT, K. and ESKENAZI, B. (2014). Prenatal exposure to dichlorodiphenyltrichloroethane and obesity at 9 years of age in the CHAMACOS study cohort. *Am. J. Epidemiol.* **179** 1312–1322.
- WEDEL, M. (2002). Concomitant variables in finite mixture models. *Stat. Neerl.* **56** 362–375.

EXTREME VALUE MODELLING OF WATER-RELATED INSURANCE CLAIMS¹

BY CHRISTIAN ROHRBECK*, EMMA F. EASTOE*, ARNOLDO FRIGESSI[†] AND JONATHAN A. TAWN*

Lancaster University and University of Oslo[†]*

This paper considers the dependence between weather events, for example, rainfall or snow-melt, and the number of water-related property insurance claims. Weather events which cause severe damages are of general interest; decision makers want to take efficient actions against them while the insurance companies want to set adequate premiums. The modelling is challenging since the underlying dynamics vary across geographical regions due to differences in topology, construction designs and climate. We develop new methodology to improve the existing models which fail to model high numbers of claims. The statistical framework is based on both mixture and extremal mixture modelling, with the latter being based on a discretized generalized Pareto distribution. Furthermore, we propose a temporal clustering algorithm and derive new covariates which lead to a better understanding of the association between claims and weather events. The modelling of the claims, conditional on the locally observed weather events, both fit the marginal distributions well and capture the spatial dependence between locations. Our methodology is applied to three cities across Norway to demonstrate its benefits.

REFERENCES

- ANDERSON, C. W. (1970). Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *J. Appl. Probab.* **7** 99–113. [MR0256441](#)
- ANDERSON, C. W. (1980). Local limit theorems for the maxima of discrete random variables. *Math. Proc. Cambridge Philos. Soc.* **88** 161–165. [MR0569642](#)
- ANDERSON, C. W., COLES, S. G. and HÜSLER, J. (1997). Maxima of Poisson-like variables and related triangular arrays. *Ann. Appl. Probab.* **7** 953–971. [MR1484793](#)
- BEHRENS, C. N., LOPES, H. F. and GAMERMAN, D. (2004). Bayesian analysis of extreme events with threshold estimation. *Stat. Model.* **4** 227–244. [MR2062102](#)
- BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. [MR0373208](#)
- BESAG, J., YORK, J. and MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43** 1–59. [MR1105822](#)
- BOLDI, M.-O. and DAVISON, A. C. (2007). A mixture model for multivariate extremes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 217–229. [MR2325273](#)
- BOTTOLO, L., CONSONNI, G., DELLAPORTAS, P. and LIJOI, A. (2003). Bayesian analysis of extreme values by mixture modeling. *Extremes* **6** 25–47. [MR2021591](#)

Key words and phrases. Extremal dependence, extremal mixture, insurance claims, mixture modelling, Poisson hurdle model, spatio-temporal modelling.

- BOTZEN, W. J. and VAN DEN BERGH, J. C. (2008). Insurance against climate change and flooding in the Netherlands: Present, future, and comparison with other countries. *Risk Anal.* **28** 413–426.
- BOTZEN, W. J. and VAN DEN BERGH, J. C. (2012). Risk attitudes to low-probability climate change risks: WTP for flood insurance. *J. Econ. Behav. Organ.* **82** 151–166.
- BROCKWELL, A. E. (2007). Universal residuals: A multivariate transformation. *Statist. Probab. Lett.* **77** 1473–1478. [MR2395595](#)
- BROOKS, S. P. and GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7** 434–455. [MR1665662](#)
- BUDDANA, A. and KOZUBOWSKI, T. J. (2014). Discrete Pareto distributions. *Econ. Qual. Control* **29** 143–156.
- CARREAU, J. and BENGIO, Y. (2009). A hybrid Pareto model for asymmetric fat-tailed data: The univariate case. *Extremes* **12** 53–76. [MR2480723](#)
- COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London. [MR1932132](#)
- COLES, S., HEFFERNAN, J. and TAWN, J. A. (1999). Dependence measures for extreme value analyses. *Extremes* **2** 339–365.
- COLES, S. G. and TAWN, J. A. (1991). Modelling extreme multivariate events. *J. Roy. Statist. Soc. Ser. B* **53** 377–392. [MR1108334](#)
- COLES, S. G. and TAWN, J. A. (1996). Modelling extremes of the areal rainfall process. *J. Roy. Statist. Soc. Ser. B* **58** 329–347. [MR1377836](#)
- DAVISON, A. C. and SMITH, R. L. (1990). Models for exceedances over high thresholds. *J. Roy. Statist. Soc. Ser. B* **52** 393–442. [MR1086795](#)
- DE CARVALHO, M. and DAVISON, A. C. (2014). Spectral density ratio models for multivariate extremes. *J. Amer. Statist. Assoc.* **109** 764–776. [MR3223748](#)
- DEPARTMENT FOR ENVIRONMENT, FOOD AND RURAL AFFAIRS (2004). Review of UK climate change indicators. Available from <http://www.ecn.ac.uk/iccuk/>.
- EASTOE, E. F. and TAWN, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 25–45. [MR2662232](#)
- EINMAHL, J. H. J., DE HAAN, L. and SINHA, A. K. (1997). Estimating the spectral measure of an extreme value distribution. *Stochastic Process. Appl.* **70** 143–171. [MR1475660](#)
- EINMAHL, J. H. J. and SEGERS, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *Ann. Statist.* **37** 2953–2989. [MR2541452](#)
- FREES, E. W. and VALDEZ, E. A. (2008). Hierarchical insurance claims modeling. *J. Amer. Statist. Assoc.* **103** 1457–1469. [MR2655723](#)
- FRIGESSI, A., HAUG, O. and RUE, H. (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes* **5** 219–235. [MR1995776](#)
- HANSON, T. E., DE CARVALHO, M. and CHEN, Y. (2017). Bernstein polynomial angular densities of multivariate extreme value distributions. *Statist. Probab. Lett.* **128** 60–66. [MR3656376](#)
- HAUG, O., DIMAKOS, X. K., VÅRDAL, J. F., ALDRIN, M. and MEZE-HAUSKEN, E. (2011). Future building water loss projections posed by climate change. *Scand. Actuar. J.* **2011** 1–20. [MR2773749](#)
- HITZ, A. (2017). Modelling of Extremes. Unpublished Oxford University PhD thesis.
- HOLMES, J. and MORIARTY, W. (1999). Application of the generalized Pareto distribution to extreme value analysis in wind engineering. *J. Wind Eng. Ind. Aerodyn.* **83** 1–10.
- JENKINS, G. J., PERRY, M. C. and PRIOR, M. J. (2008). The climate of the United Kingdom and recent trends. Met Office Hadley Centre, Exeter, UK.
- KLUGMAN, S. A., PANJER, H. H. and WILLMOT, G. E. (2012). *Loss Models: From Data to Decisions*, 4th ed. Wiley, Hoboken, NJ. [MR3222004](#)
- KNABB, R. D., RHOME, J. R. and BROWN, D. P. (2005). Tropical Cyclone Report: Hurricane Katrina: 23–30 August 2005. Technical report, National Hurricane Center.

- KUBILAY, A., DEROME, D., BLOCKEN, B. and CARMELIET, J. (2013). CFD simulation and validation of wind-driven rain on a building facade with an Eulerian multiphase model. *Build. Environ.* **61** 69–81.
- LEDFORD, A. W. and TAWN, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika* **83** 169–187. [MR1399163](#)
- LI, Y., CAI, W. and CAMPBELL, E. (2005). Statistical modeling of extreme rainfall in southwest Western Australia. *J. Climate* **18** 852–863.
- MACDONALD, A., SCARROTT, C. J., LEE, D., DARLOW, B., REALE, M. and RUSSELL, G. (2011). A flexible extreme value mixture model. *Comput. Statist. Data Anal.* **55** 2137–2157. [MR2785120](#)
- MILLS, E. (2005). Insurance industry in a climate of change. *Science* **309** 1040–1044.
- NELSEN, R. B. (2006). *An Introduction to Copulas*, 2nd ed. Springer, New York. [MR2197664](#)
- PICKANDS, J. III (1971). The two-dimensional Poisson process and extremal processes. *J. Appl. Probab.* **8** 745–756. [MR0295453](#)
- PRIETO, F., GÓMEZ-DÉNIZ, E. and SARABIA, J. M. (2014). Modelling road accident blackspots data with the discrete generalized Pareto distribution. *Accident Anal. Prev.* **71** 38–49.
- RESNICK, S. I. (2013). *Extreme Values, Regular Variation and Point Processes*. Springer, New York. [MR2364939](#)
- RUSSELL, B. T., COOLEY, D. S., PORTER, W. C., REICH, B. J. and HEALD, C. L. (2016). Data mining to investigate the meteorological drivers for extreme ground level ozone events. *Ann. Appl. Stat.* **10** 1673–1698. [MR3553240](#)
- SANDERS, C. H. and PHILLIPSON, M. C. (2003). UK adaption strategy and technical measures: The impacts of climate change on buildings. *Build. Res. Inf.* **31** 210–221.
- SCHEEL, I., FERKINGSTAD, E., FRIGESSI, A., HAUG, O., HINNERICHSEN, M. and MEZEHAUSKEN, E. (2013). A Bayesian hierarchical model with spatial variable selection: The effect of weather on insurance claims. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 85–100. [MR3042312](#)
- SCHUSTER, S. S., BLONG, R. J. and MCANENEY, K. J. (2006). Relationship between radar-derived hail kinetic energy and damage to insured buildings for severe hailstorms in Eastern Australia. *Atmos. Res.* **81** 215–235.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. [MR0468014](#)
- SHIMURA, T. (2012). Discretization of distributions in the maximum domain of attraction. *Extremes* **15** 299–317. [MR2969061](#)
- SMITH, J. (1985). Diagnostic checks of non-standard time series models. *J. Forecast.* **4** 283–291.
- SMITH, R. L. and GOODMAN, D. J. (2000). Bayesian risk analysis. In *Extremes and Integrated Risk Management* (P. Embrechts, ed.) 235–251. Risk Books, London.
- 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](#)
- YIP, K. C. and YAU, K. K. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance Math. Econom.* **36** 153–163.

TIME-VARYING EXTREME VALUE DEPENDENCE WITH APPLICATION TO LEADING EUROPEAN STOCK MARKETS¹

BY DANIELA CASTRO-CAMILO, MIGUEL DE CARVALHO AND
 JENNIFER WADSWORTH

*King Abdullah University of Science and Technology,
 University of Edinburgh, and Lancaster University*

Extremal dependence between international stock markets is of particular interest in today's global financial landscape. However, previous studies have shown this dependence is not necessarily stationary over time. We concern ourselves with modeling extreme value dependence when that dependence is changing over time, or other suitable covariate. Working within a framework of asymptotic dependence, we introduce a regression model for the angular density of a bivariate extreme value distribution that allows us to assess how extremal dependence evolves over a covariate. We apply the proposed model to assess the dynamics governing extremal dependence of some leading European stock markets over the last three decades, and find evidence of an increase in extremal dependence over recent years.

REFERENCES

- ACAR, E. F., CRAIU, R. V. and YAO, F. (2011). Dependence calibration in conditional copulas: A nonparametric approach. *Biometrics* **67** 445–453. [MR2829013](#)
- BEIRLANT, J., GOEGBEUR, Y., TEUGELS, J. and SEGERS, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, Chichester. [MR2108013](#)
- BROOKS, R. and DEL NEGRO, M. (2004). The rise in comovement across national stock markets: Market integration or IT bubble? *J. Empir. Finance* **11** 659–680.
- BROOKS, R. and DEL NEGRO, M. (2005). Country versus region effects in international stock returns. *J. Portf. Manag.* **31** 67–72.
- BÜTTNER, D. and HAYO, B. (2011). Determinants of European stock market integration. *Econ. Syst.* **35** 574–585.
- CASTRO-CAMILO, D. and DE CARVALHO, M. (2017). Spectral density regression for bivariate extremes. *Stoch. Environ. Res. Risk Assess.* **31** 1603–1613.
- CASTRO-CAMILO, D., DE CARVALHO, M. and WADSWORTH, J. (2018). Supplement to “Time-varying extreme value dependence with application to leading European stock markets.” DOI:[10.1214/17-AOAS1089SUPP](https://doi.org/10.1214/17-AOAS1089SUPP).
- CHAVEZ-DEMOULIN, V. and DAVISON, A. C. (2005). Generalized additive modelling of sample extremes. *J. Roy. Statist. Soc. Ser. C* **54** 207–222. [MR2134607](#)
- CHEN, S. X. (1999). Beta kernel estimators for density functions. *Comput. Statist. Data Anal.* **31** 131–145. [MR1718494](#)
- CLINE, D. B. H. (1988). Admissible kernel estimators of a multivariate density. *Ann. Statist.* **16** 1421–1427. [MR0964931](#)

Key words and phrases. Angular measure, bivariate extreme values, European stock market integration, risk, statistics of extremes.

- COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London. [MR1932132](#)
- COLES, S. G. and TAWN, J. A. (1991). Modelling extreme multivariate events. *J. Roy. Statist. Soc. Ser. B* **53** 377–392. [MR1108334](#)
- DASGUPTA, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer, New York. [MR2664452](#)
- DAVISON, A. C. and SMITH, R. L. (1990). Models for exceedances over high thresholds. *J. Roy. Statist. Soc. Ser. B* **52** 393–442. [MR1086795](#)
- DE CARVALHO, M. (2017). Statistics of extremes: Challenges and opportunities. In *Extreme Events in Finance. Wiley Handb. Finance Eng. Econom.* 195–213. Wiley, Hoboken, NJ. [MR3617263](#)
- DE CARVALHO, M. and DAVISON, A. C. (2014). Spectral density ratio models for multivariate extremes. *J. Amer. Statist. Assoc.* **109** 764–776. [MR3223748](#)
- DE CARVALHO, M., OUMOW, B., SEGERS, J. and WARCHOL, M. (2013). A Euclidean likelihood estimator for bivariate tail dependence. *Comm. Statist. Theory Methods* **42** 1176–1192. [MR3031275](#)
- DE HAAN, L. and RESNICK, S. I. (1977). Limit theory for multivariate sample extremes. *Z. Wahrsch. Verw. Gebiete* **40** 317–377.
- EASTOE, E. F. (2009). A hierarchical model for non-stationary multivariate extremes: A case study of surface-level ozone and NO_x data in the UK. *Environmetrics* **20** 428–444. [MR2834808](#)
- EASTOE, E. F. and TAWN, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *J. R. Stat. Soc. Ser. C* **58** 25–45. [MR2662232](#)
- ENGLE, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50** 987–1007. [MR0666121](#)
- ESCOBAR-BACH, M., GOEGBEUR, J. and GUILLOU, A. (2016). Local robust estimation of the Pickands dependence function. Submitted.
- FERMANIAN, J.-D. and WEGKAMP, M. H. (2012). Time-dependent copulas. *J. Multivariate Anal.* **110** 19–29. [MR2927507](#)
- FORBES, K. and RIGOBON, R. (2002). No contagion, only interdependence: Measuring stock market comovements. *J. Finance* **57** 2223–2261.
- FRATZSCHER, M. (2002). Financial market integration in Europe: On the effects of EMU on stock markets. *Int. J. Fin. Econ.* **7** 165–193.
- GEENENS, G. (2014). Probit transformation for kernel density estimation on the unit interval. *J. Amer. Statist. Assoc.* **109** 346–358. [MR3180568](#)
- GUDENDORF, G. and SEGERS, J. (2010). Extreme-value copulas. In *Copula Theory and Its Applications. Lect. Notes Stat. Proc.* **198** 127–145. Springer, Heidelberg. [MR3051266](#)
- HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press, Cambridge. [MR1161622](#)
- HARDOUVELIS, G. A., MALLIAROPULOS, D. and PRIESTLEY, R. (2006). EMU and European stock market integration. *J. Bus.* **79** 365–392.
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York. [MR1851606](#)
- HEFFERNAN, J. E. (2000). A directory of coefficients of tail dependence. *Extremes* **3** 279–290. [MR1856201](#)
- HEFFERNAN, J. E. and TAWN, J. A. (2004). A conditional approach for multivariate extreme values. *J. R. Stat. Soc. Ser. B* **66** 497–546. [MR2088289](#)
- HUSER, R. and GENTON, M. G. (2016). Non-stationary dependence structures for spatial extremes. *J. Agric. Biol. Environ. Stat.* **21** 470–491. [MR3542082](#)
- JAMES, H. (2012). *Making the European Monetary Union*. Harvard Univ. Press, Cambridge MA.
- JONATHAN, P., EWANS, K. and RANDELL, D. (2014). Non-stationary conditional extremes of northern North Sea storm characteristics. *Environmetrics* **25** 172–188. [MR3200308](#)

- JONES, M. C. and HENDERSON, D. A. (2007). Kernel-type density estimation on the unit interval. *Biometrika* **94** 977–984. [MR2416803](#)
- KAROLYI, G. A. and STULZ, R. M. (1996). Why do markets move together? An investigation of US–Japan stock return comovements. *J. Finance* **51** 951–986.
- KIM, S. J., MOSHIRIAN, F. and WU, E. (2005). Dynamic stock market integration driven by the European monetary union: An empirical analysis. *J. Bank. Fin.* **29** 2475–2502.
- KIM, S., SHEPHARD, N. and CHIB, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Rev. Econ. Stud.* **65** 361–393.
- KING, M., SENTANA, E. and SUSHIL, W. (1994). Volatility and links between national stock markets. *Econometrica* **62** 901–933.
- KOENKER, R. (2005). *Quantile Regression*. Cambridge Univ. Press, Cambridge. [MR2268657](#)
- LANE, P. R. (2012). The European sovereign debt crisis. *J. Econ. Persp.* **26** 49–67.
- LONGIN, F. and SOLNIK, B. (1995). Is the correlation in international equity returns constant: 1960–1990? *J. Int. Money Financ.* **14** 3–26.
- LONGIN, F. and SOLNIK, B. (2001). Extreme correlation of international equity market. *J. Finance* **56** 649–676.
- MHALLA, L., CHAVEZ-DEMOULIN, V. and NAVEAU, P. (2017). Non-linear models for extremal dependence. *J. Multivariate Anal.* **159** 49–66. [MR3668547](#)
- NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.
- NATURAL ENVIRONMENT RESEARCH COUNCIL (1975). *Flood Studies Report*. Natural Environment Research Council, London.
- PATTON, A. J. (2006). Modelling asymmetric exchange rate dependence. *Internat. Econom. Rev.* **47** 527–556. [MR2216591](#)
- PICKANDS, J. III (1981). Multivariate extreme value distributions. In *Proceedings of the 43rd Session of the International Statistical Institute, Vol. 2 (Buenos Aires, 1981)* **49** 859–878, 894–902. [MR0820979](#)
- POLANSKY, A. M. (2001). Bandwidth selection for the smoothed bootstrap percentile method. *Comput. Statist. Data Anal.* **36** 333–349. [MR1836205](#)
- POON, S.-H., ROCKINGER, M. and TAWN, J. A. (2003). Modelling extreme-value dependence in international stock markets. *Statist. Sinica* **13** 929–953.
- POON, S.-H., ROCKINGER, M. and TAWN, J. (2004). Extreme-value dependence in financial markets: Diagnostics, models and financial implications. *Rev. Financ. Stud.* **17** 581–610.
- RESNICK, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, New York. [MR2271424](#)
- RUA, A. and NUNES, L. C. (2009). International comovement of stock market returns: A wavelet analysis. *J. Empir. Finance* **16** 632–639.
- SILVERMAN, B. W. and YOUNG, G. A. (1987). The bootstrap: To smooth or not to smooth? *Biometrika* **74** 469–479. [MR0909352](#)
- SUN, Y. and GENTON, M. G. (2011). Functional boxplots. *J. Comput. Graph. Statist.* **20** 316–334. [MR2847798](#)
- TAWN, J. A. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika* **75** 397–415. [MR0967580](#)
- TAWN, J. A. (1990). Modelling multivariate extreme value distributions. *Biometrika* **77** 245–253.
- VERAVERBEKE, N., OMELKA, M. and GIJBELS, I. (2011). Estimation of a conditional copula and association measures. *Scand. J. Stat.* **38** 766–780. [MR2859749](#)
- WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman & Hall, London. [MR1319818](#)
- WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372. [MR0185765](#)

AUTOMATED THRESHOLD SELECTION FOR EXTREME VALUE ANALYSIS VIA ORDERED GOODNESS-OF-FIT TESTS WITH ADJUSTMENT FOR FALSE DISCOVERY RATE

BY BRIAN BADER^{*,†}, JUN YAN^{†,1} AND XUEBIN ZHANG[‡]

KPMG LLP^{}, University of Connecticut[†], and Environment and Climate Change Canada[‡]*

Threshold selection is a critical issue for extreme value analysis with threshold-based approaches. Under suitable conditions, exceedances over a high threshold have been shown to follow the generalized Pareto distribution (GPD) asymptotically. In practice, however, the threshold must be chosen. If the chosen threshold is too low, the GPD approximation may not hold and bias can occur. If the threshold is chosen too high, reduced sample size increases the variance of parameter estimates. To process batch analyses, commonly used selection methods such as graphical diagnostics are subjective and cannot be automated. We develop an efficient technique to evaluate and apply the Anderson–Darling test to the sample of exceedances above a fixed threshold. In order to automate threshold selection, this test is used in conjunction with a recently developed stopping rule that controls the false discovery rate in ordered hypothesis testing. Previous attempts in this setting do not account for the issue of ordered multiple testing. The performance of the method is assessed in a large scale simulation study that mimics practical return level estimation. This procedure was repeated at hundreds of sites in the western US to generate return level maps of extreme precipitation.

REFERENCES

- BADER, B. and YAN, J. (2015). *eva*: Extreme value analysis with goodness-of-fit testing. R package version 0.1.2.
- BADER, B., YAN, J. and ZHANG, X. (2018). Supplement to “Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate.” DOI:10.1214/17-AOAS1092SUPP.
- BALKEMA, A. A. and DE HAAN, L. (1974). Residual life time at great age. *Ann. Probab.* **2** 792–804. MR0359049
- BENJAMINI, Y. (2010a). Discovering the false discovery rate. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 405–416.
- BENJAMINI, Y. (2010b). Simultaneous and selective inference: Current successes and future challenges. *Biom. J.* **52** 708–721. MR2758547
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **57** 289–300.
- BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245
- BLANCHARD, G. and ROQUAIN, É. (2009). Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.* **10** 2837–2871.

Key words and phrases. Batch analysis, exceedance diagnostic, specification test, stopping rule.

- BLANCHET, J. and LEHNING, M. (2010). Mapping snow depth return levels: Smooth spatial modeling versus station interpolation. *Hydrology and Earth System Sciences* **14** 2527–2544.
- CAEIRO, F. and GOMES, M. I. (2015). Threshold selection in extreme value analysis. In *Extreme Value Modeling and Risk Analysis: Methods and Applications* (D. K. Dey and J. Yan, eds.) 69–82. CRC Press, Boca Raton.
- CAIRES, S. (2009). A Comparative Simulation Study of the Annual Maxima and the Peaks-over-Threshold Methods Technical report, SBW-Belastingen: Subproject “Statistics”. Deltares Report 1200264-002.
- CHENG, R. C. H. and STEPHENS, M. A. (1989). A goodness-of-fit test using Moran’s statistic with estimated parameters. *Biometrika* **76** 385–392. [MR1016030](#)
- CHOULAKIAN, V. and STEPHENS, M. A. (2001). Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics* **43** 478–484.
- COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, 1st ed. Springer, Berlin.
- DANIELSSON, J., DE HAAN, L., PENG, L. and DE VRIES, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *J. Multivariate Anal.* **76** 226–248.
- DAVISON, A. C. and SMITH, R. L. (1990). Models for exceedances over high thresholds. *J. Roy. Statist. Soc. Ser. B* **52** 393–442. [MR1086795](#)
- DEIDDA, R. and PULIGA, M. (2006). Sensitivity of goodness-of-fit statistics to rainfall data rounding off. *Physics and Chemistry of the Earth, Parts A/B/C* **31** 1240–1251.
- DEIDDA, R. and PULIGA, M. (2009). Performances of some parameter estimators of the generalized Pareto distribution over rounded-off samples. *Physics and Chemistry of the Earth, Parts A/B/C* **34** 626–634.
- DEY, D. K. and YAN, J., eds. (2015). *Extreme Value Modeling and Risk Analysis: Methods and Applications*. CRC Press, Boca Raton.
- DREES, H., DE HAAN, L. and RESNICK, S. (2000). How to make a hill plot. *Ann. Statist.* **28** 254–274.
- DUMOUCHEL, W. H. (1983). Estimating the stable index α in order to measure tail thickness: A critique. *Ann. Statist.* **11** 1019–1031. [MR0865342](#)
- DUPUIS, D. J. (1999). Exceedances over high thresholds: A guide to threshold selection. *Extremes* **1** 251–261.
- FAWCETT, L. and WALSHAW, D. (2007). Improved estimation for temporally clustered extremes. *Environmetrics* **18** 173–188.
- FERREIRA, A., DE HAAN, L. and PENG, L. (2003). On optimising the estimation of high quantiles of a probability distribution. *Statistics* **37** 401–434. [MR2006789](#)
- FERRO, C. A. and SEGERS, J. (2003). Inference for clusters of extreme values. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 545–556.
- FISHER, R. A. and TIPPETT, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society* **24** 180–190. Cambridge Univ. Press, Cambridge.
- G’SELL, M. G., WAGER, S., CHOULDECHOVA, A. and TIBSHIRANI, R. (2016). Sequential selection procedures and false discovery rate control. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 423–444. [MR3454203](#)
- GOEGBEUR, Y., BEIRLANT, J. and DE WET, T. (2008). Linking Pareto-tail kernel goodness-of-fit statistics with tail index at optimal threshold and second order estimation. *REVSTAT* **6** 51–69.
- HOLDEN, L. and HAUG, O. (2009). A Multidimensional Mixture Model for Unsupervised Tail Estimation. NR-notat SAMBA/09/09. pp 29.
- JACKSON, O. (1967). An analysis of departures from the exponential distribution. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 540–549.
- KATZ, R. W., PARLANGE, M. B. and NAVEAU, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources* **25** 1287–1304.

- KHARIN, V. V., ZWIERS, F. W., ZHANG, X. and HEGERL, G. C. (2007). Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. *J. Climate* **20** 1419–1444.
- KHARIN, V. V., ZWIERS, F., ZHANG, X. and WEHNER, M. (2013). Changes in temperature and precipitation extremes in the CMIP5 ensemble. *Climatic Change* **119** 345–357.
- LANGOUSIS, A., MAMALAKIS, A., PULIGA, M. and DEIDDA, R. (2016). Threshold detection for the generalized Pareto distribution: Review of representative methods and application to the NOAA NCDC daily rainfall database. *Water Resources Research* **52** 2659–2681.
- LATELTIN, O. and BONNARD, C. (1999). Hazard Assessment and Land-Use Planning in Switzerland for Snow Avalanches, Floods and Landslides. Technical report, World Meteorological Organization.
- LEADBETTER, M. R., WEISSMAN, I., DE HAAN, L. and ROOTZÉN, H. (1989). On clustering of high values in statistically stationary series. *Proc. 4th Int. Meet. Statistical Climatology* **16** 217–222.
- LEWIS, P. A. W. (1965). Some results on tests for Poisson processes. *Biometrika* **52** 67–77. [MR0207107](#)
- MACDONALD, A., SCARROTT, C. J., LEE, D., DARLOW, B., REALE, M. and RUSSELL, G. (2011). A flexible extreme value mixture model. *Comput. Statist. Data Anal.* **55** 2137–2157.
- MENNE, M. J., DURRE, I., VOSE, R. S., GLEASON, B. E. and HOUSTON, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology* **29** 897–910.
- MORAN, P. A. P. (1953). The random division of an interval—part II. *J. Roy. Statist. Soc. Ser. B* **15** 77–80.
- NADARAJAH, S. and ELJABRI, S. (2013). The Kumaraswamy GP distribution. *J. Data Sci.* **11** 739–766.
- NAVEAU, P., HUSER, R., RIBEREAU, P. and HANNART, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research* **52** 2753–2769.
- NORTHROP, P. J., ATTALIDES, N. and JONATHAN, P. (2017). Cross-validators extreme value threshold selection and uncertainty with application to ocean storm severity. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 93–120.
- NORTHROP, P. J. and COLEMAN, C. L. (2014). Improved threshold diagnostic plots for extreme value analyses. *Extremes* **17** 289–303.
- NORTHROP, P. J. and JONATHAN, P. (2011). Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics* **22** 799–809. [MR2861046](#)
- PAPALEXIOU, S. M. and KOUTSOYIANNIS, D. (2013). Battle of extreme value distributions: A global survey on extreme daily rainfall. *Water Resources Research* **49** 187–201.
- PAPASTATHOPOULOS, I. and TAWN, J. A. (2013). Extended generalised Pareto models for tail estimation. *J. Statist. Plann. Inference* **143** 131–143. [MR2969016](#)
- PICKANDS, J. III (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3** 119–131. [MR0423667](#)
- RAOULT, J.-P. and WORMS, R. (2003). Rate of convergence for the generalized Pareto approximation of the excesses. *Adv. in Appl. Probab.* **35** 1007–1027.
- ROTH, M., JONGBLOED, G. and BUIHAND, T. A. (2016). Threshold selection for regional peaks-over-threshold data. *J. Appl. Stat.* **43** 1291–1309. [MR3473771](#)
- ROTH, M., BUIHAND, T. A., JONGBLOED, G., KLEIN TANK, A. M. G. and VAN ZANTEN, J. H. (2012). A regional peaks-over-threshold model in a nonstationary climate. *Water Resources Research* **48**.
- SCARROTT, C. and MACDONALD, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT* **10** 33–60.

- SERINALDI, F. and KILSBY, C. G. (2014). Rainfall extremes: Toward reconciliation after the battle of distributions. *Water Resources Research* **50** 336–352.
- SOUTHWORTH, H. and HEFFERNAN, J. E. (2012). texmex: Threshold Exceedences and Multivariate Extremes. R package version 1.3.
- THOMPSON, P., CAI, Y., REEVE, D. and STANDER, J. (2009). Automated threshold selection methods for extreme wave analysis. *Coastal Engineering* **56** 1013–1021.
- WADSWORTH, J. L. (2016). Exploiting structure of maximum likelihood estimators for extreme value threshold selection. *Technometrics* **58** 116–126.
- WADSWORTH, J. L. and TAWN, J. A. (2012). Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 543–567.
- WANG, Q. J. (1991). The POT model described by the generalized Pareto distribution with Poisson arrival rate. *J. Hydrol.* **129** 263–280.
- WONG, T. S. T. and LI, W. K. (2006). A note on the estimation of extreme value distributions using maximum product of spacings. In *Time Series and Related Topics* 272–283. IMS.

DESIGN OF VACCINE TRIALS DURING OUTBREAKS WITH AND WITHOUT A DELAYED VACCINATION COMPARATOR¹

BY NATALIE E. DEAN^{*}, M. ELIZABETH HALLORAN^{†,‡} AND
 IRA M. LONGINI^{*}

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

Conducting vaccine efficacy trials during outbreaks of emerging pathogens poses particular challenges. The “Ebola ça suffit” trial in Guinea used a novel ring vaccination cluster randomized design to target populations at highest risk of infection. Another key feature of the trial was the use of a delayed vaccination arm as a comparator, in which clusters were randomized to immediate vaccination or vaccination 21 days later. This approach, chosen to improve ethical acceptability of the trial, complicates the statistical analysis as participants in the comparison arm are eventually protected by vaccine. Furthermore, for infectious diseases, we observe time of illness onset and not time of infection, and we may not know the time required for the vaccinee to develop a protective immune response. As a result, including events observed shortly after vaccination may bias the per protocol estimate of vaccine efficacy. We provide a framework for approximating the bias and power of any given analysis period as functions of the background infection hazard rate, disease incubation period, and vaccine immune response. We use this framework to provide recommendations for designing standard vaccine efficacy trials and trials with a delayed vaccination comparator. Briefly, narrower analysis periods within the correct window can minimize or eliminate bias but may suffer from reduced power. Designs should be reasonably robust to misspecification of the incubation period and time to develop a vaccine immune response.

REFERENCES

- BELLAN, S. E., PULLIAM, J. R. C., PEARSON, C. A. B., CHAMPREDON, D., FOX, S. J., SKRIP, L., GALVANI, A. P., GAMBHIR, M., LOPMAN, B. A., PORCO, T. C., MEYERS, L. A. and DUSHOFF, J. (2015). Statistical power and validity of Ebola vaccine trials in Sierra Leone: A simulation study of trial design and analysis. *Lancet, Infect. Dis.* **15** 703–710.
- EBOLA ÇA SUFFIT RING VACCINATION TRIAL CONSORTIUM (2015). The ring vaccination trial: A novel cluster randomised controlled trial design to evaluate vaccine efficacy and effectiveness during outbreaks, with special reference to Ebola. *Br. Med. J.* **351** h3740.
- CAMACHO, A., EGGO, R. M., FUNK, S., WATSON, C. H., KUCHARSKI, A. J. and EDMUNDS, W. J. (2015). Estimating the probability of demonstrating vaccine efficacy in the declining Ebola epidemic: A Bayesian modelling approach. *BMJ Open* **5** e009346.
- DEAN, N., HALLORAN, M. E. and LONGINI, I. M. (2018). Supplement to “Design of vaccine trials during outbreaks with and without a delayed vaccination comparator.” DOI:10.1214/17-AOAS1095SUPP.

Key words and phrases. Vaccine trial, infectious diseases, per protocol.

- GILBERT, P. B., BERGER, J. O., STABLEIN, D., BECKER, S., ESSEX, M., HAMMER, S. M., KIM, J. H. and DEGRUTTOLA, V. G. (2011). Statistical interpretation of the RV144 HIV vaccine efficacy trial in Thailand: A case study for statistical issues in efficacy trials. *J. Infect. Dis.* **203** 969–975.
- HALLORAN, M. E., LONGINI, I. M. JR. and STRUCHINER, C. J. (2010). *Design and Analysis of Vaccine Studies*. Springer, New York.
- HALLORAN, M. E., HABER, M., LONGINI, I. M. and STRUCHINER, C. J. (1991). Direct and indirect effects in vaccine efficacy and effectiveness. *Am. J. Epidemiol.* **133** 323–331.
- HENAO-RESTREPO, A. M., LONGINI, I. M., EGGER, M., DEAN, N. E., EDMUNDS, W. J., CAMACHO, A., CARROLL, M. W., DOUMBIA, M., DRAGUEZ, B., DURAFFOUR, S., ENWERE, G., GRAIS, R., GUNTHER, S., HOSSMANN, S., KONDE, M. K., KONE, S., KUISMA, E., LEVINE, M. M., MANDAL, S., NORHEIM, G., RIVEROS, X., SOUMAH, A., TRELLE, S., VICARI, A. S., WATSON, C. H., KEITA, S., KIENY, M.-P. P. and RØTTINGEN, J. A. (2015). Efficacy and effectiveness of an rVSV-vectored vaccine expressing Ebola surface glycoprotein: Interim results from the Guinea ring vaccination cluster-randomised trial. *Lancet* **386** 857–866.
- HENAO-RESTREPO, A. M., CAMACHO, A., LONGINI, I. M., WATSON, C. H., EDMUNDS, W. J., EGGER, M., CARROLL, M. W., DEAN, N. E., DATTA, I., DOUMBIA, M., DRAGUEZ, B., DURAFFOUR, S., ENWERE, G., GRAIS, R., GUNTHER, S., GSELL, P.-S., HOSSMANN, S., WATTLE, S. V., KONDE, M. K., KEITA, S., KONE, S., KUISMA, E., LEVINE, M. M., MANDAL, S., MAUGET, T., NORHEIM, G., RIVEROS, X., SOUMAH, A., TRELLE, S., VICARI, A. S., RØTTINGEN, J. A. and KIENY, M.-P. (2017). Efficacy and effectiveness of an rVSV-vectored vaccine preventing Ebola virus disease: Final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola ça suffit!). *Lancet* **389** 505–518.
- HERNÁN, M. A. (2013). The hazards of hazard ratios. *Epidemiology* **21** 13–15.
- HORNE, A. D., LACHENBRUCH, P. A. and GOLDENTHAL, K. L. (2001). Intent-to-treat analysis and preventive vaccine efficacy. *Vaccine* **19** 319–326.
- HUSSEY, M. A. and HUGHES, J. P. (2007). Design and analysis of stepped wedge cluster randomised trials. *Contemp. Clin. Trials* **28** 182–191.
- KIENY, M. P. and SALAMA, P. (2017). WHO R&D blueprint: A global coordination mechanism for R&D preparedness. *Lancet* **389** 2469–2470.
- RTSS CLINICAL TRIALS PARTNERSHIP (2015). Efficacy and safety of RTS, S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: Final results of a phase 3, individually randomised, controlled trial. *Lancet* **386** 31–45.
- RIDOUT, M. S., DEMÉTRIO, C. G. and FIRTH, D. (1999). Estimating intraclass correlation for binary data. *Biometrics* **55** 137–148.
- ROSNER, B. (2010). *Fundamentals of Biostatistics*, 7th ed. Cengage Learning, Boston.
- R CORE TEAM (2015). *A Language and Environment for Statistical Computing*. Available at <http://www.r-project.org/>.
- THERNEAU, T. M. (2015). Package ‘survival’: Survival analysis. R package. Available at <http://cran.r-project.org/web/packages/survival/index.html>.
- XU, R. and O’QUIGLEY, J. (2000). Estimating average regression effect under non-proportional hazards. *Biostatistics* **1** 423–439.
- ZUCKER, D. M. and LAKATOS, E. (1990). Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* **77** 853–864.
- MR1086695

TWO-LEVEL STRUCTURAL SPARSITY REGULARIZATION FOR IDENTIFYING LATTICES AND DEFECTS IN NOISY IMAGES

BY XIN LI^{*,1}, ALEX BELIANINOV^{†,2}, ONDREJ DYCK^{†,2},
STEPHEN JESSE^{†,2} AND CHIWOO PARK^{*,1}

Florida State University and Oak Ridge National Laboratory†*

This paper presents a regularized regression model with a two-level structural sparsity penalty applied to locate individual atoms in a noisy scanning transmission electron microscopy image (STEM). In crystals, the locations of atoms is symmetric, condensed into a few lattice groups. Therefore, by identifying the underlying lattice in a given image, individual atoms can be accurately located. We propose to formulate the identification of the lattice groups as a sparse group selection problem. Furthermore, real atomic scale images contain defects and vacancies, so atomic identification based solely on a lattice group may result in false positives and false negatives. To minimize error, model includes an individual sparsity regularization in addition to the group sparsity for a within-group selection, which results in a regression model with a two-level sparsity regularization. We propose a modification of the group orthogonal matching pursuit (gOMP) algorithm with a thresholding step to solve the atom finding problem. The convergence and statistical analyses of the proposed algorithm are presented. The proposed algorithm is also evaluated through numerical experiments with simulated images. The applicability of the algorithm on determination of atom structures and identification of imaging distortions and atomic defects was demonstrated using three real STEM images. We believe this is an important step toward automatic phase identification and assignment with the advent of genomic databases for materials.

REFERENCES

- ADLER, S. B. (2004). Factors governing oxygen reduction in solid oxide fuel cell cathodes. *Chem. Rev.* **104** 4791–4843.
- BAY, H., ESS, A., TUYTELAARS, T. and VAN GOOL, L. (2008). Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **110** 346–359.
- BELIANINOV, A., HE, Q., KRAVCHENKO, M., JESSE, S., BORISEVICH, A. and KALININ, S. V. (2015). Identification of phases, symmetries and defects through local crystallography. *Nat. Commun.* **6** 7801.
- BORISEVICH, A. Y., CHANG, H. J., HUIJIBEN, M., OXLEY, M. P., OKAMOTO, S., NIRANJAN, M. K., BURTON, J., TSYMBAL, E., CHU, Y.-H., YU, P. et al. (2010a). Suppression of octahedral tilts and associated changes in electronic properties at epitaxial oxide heterostructure interfaces. *Phys. Rev. Lett.* **105** 087204.

Key words and phrases. Sparse regression, structural sparsity, lattice group, structural evaluation of materials, image data analysis.

- BORISEVICH, A., OVCHINNIKOV, O. S., CHANG, H. J., OXLEY, M. P., YU, P., SEIDEL, J., ELISEEV, E. A., MOROZOVSKA, A. N., RAMESH, R., PENNYCOOK, S. J. et al. (2010b). Mapping octahedral tilts and polarization across a domain wall in BiFeO₃ from Z-contrast scanning transmission electron microscopy image atomic column shape analysis. *ACS Nano* **4** 6071–6079.
- BRADZIL, J. F. (2010). Acrylonitrile. In *Kirk–Othmer Encyclopedia of Chemical Technology*. Wiley, New York.
- BRIGHT, D. S. and STEEL, E. B. (1987). Two-dimensional top hat filter for extracting spots and spheres from digital images. *J. Microsc.* **146** 191–200.
- CHANG, H. J., KALININ, S. V., MOROZOVSKA, A. N., HUIJBEN, M., CHU, Y.-H., YU, P., RAMESH, R., ELISEEV, E. A., SVECHNIKOV, G. S., PENNYCOOK, S. J. et al. (2011). Atomically resolved mapping of polarization and electric fields across ferroelectric/oxide interfaces by Z-contrast imaging. *Adv. Mater.* **23** 2474–2479.
- CHATTERJEE, S., STEINHAEUSER, K., BANERJEE, A., CHATTERJEE, S. and GANGULY, A. (2012). Sparse group lasso: Consistency and climate applications. In *Proceedings of the 2012 SIAM International Conference on Data Mining* 47–58. SIAM, Philadelphia, PA.
- DIMA, A., BHASKARLA, S., BECKER, C., BRADY, M., CAMPBELL, C., DESSAUW, P., HANISCH, R., KATTNER, U., KROENLEIN, K., NEWROCK, M. et al. (2016). Informatics infrastructure for the materials genome initiative. *JOM* **68** 2053–2064.
- GENOVESE, C. R., JIN, J., WASSERMAN, L. and YAO, Z. (2012). A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.* **13** 2107–2143.
- HE, J., BORISEVICH, A., KALININ, S. V., PENNYCOOK, S. J. and PANTELIDES, S. T. (2010). Control of octahedral tilts and magnetic properties of perovskite oxide heterostructures by substrate symmetry. *Phys. Rev. Lett.* **105** 227203.
- HE, Q., WOO, J., BELIANINOV, A., GULIANTS, V. V. and BORISEVICH, A. Y. (2015). Better catalysts through microscopy: Mesoscale M1/M2 intergrowth in molybdenum–vanadium based complex oxide catalysts for propane ammoxidation. *ACS Nano* **9** 3470–3478.
- HUANG, J., ZHANG, T. and METAXAS, D. (2011). Learning with structured sparsity. *J. Mach. Learn. Res.* **12** 3371–3412. [MR2877603](#)
- HUGHES, J., FRICKS, J. and HANCOCK, W. (2010). Likelihood inference for particle location in fluorescence microscopy. *Ann. Appl. Stat.* **4** 830–848. [MR2758423](#)
- JENATTON, R., AUDIBERT, J.-Y. and BACH, F. (2011). Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.* **12** 2777–2824.
- JIA, C.-L., NAGARAJAN, V., HE, J.-Q., HOUBEN, L., ZHAO, T., RAMESH, R., URBAN, K. and WASER, R. (2007). Unit-cell scale mapping of ferroelectricity and tetragonality in epitaxial ultrathin ferroelectric films. *Nat. Mater.* **6** 64–69.
- JIA, C., MI, S., FALEY, M., POPPE, U., SCHUBERT, J. and URBAN, K. (2009). Oxygen octahedron reconstruction in the SrTiO₃/LaAlO₃ heterointerfaces investigated using aberration-corrected ultrahigh-resolution transmission electron microscopy. *Phys. Rev. B* **79** 081405.
- JIA, C.-L., URBAN, K. W., ALEXE, M., HESSE, D. and VREJOIU, I. (2011). Direct observation of continuous electric dipole rotation in flux-closure domains in ferroelectric Pb(Zr,Ti)O₃. *Science* **331** 1420–1423.
- KIM, Y.-M., HE, J., BIEGALSKI, M. D., AMBAYE, H., LAUTER, V., CHRISTEN, H. M., PANTELIDES, S. T., PENNYCOOK, S. J., KALININ, S. V. and BORISEVICH, A. Y. (2012). Probing oxygen vacancy concentration and homogeneity in solid-oxide fuel-cell cathode materials on the subunit-cell level. *Nat. Mater.* **11** 888–894.
- KIM, Y.-M., KUMAR, A., HATT, A., MOROZOVSKA, A. N., TSELEV, A., BIEGALSKI, M. D., IVANOV, I., ELISEEV, E. A., PENNYCOOK, S. J., RONDINELLI, J. M. et al. (2013). Interplay of octahedral tilts and polar order in BiFeO₃ films. *Adv. Mater.* **25** 2497–2504.
- KIM, Y.-M., MOROZOVSKA, A., ELISEEV, E., OXLEY, M. P., MISHRA, R., SELBACH, S. M., GRANDE, T., PANTELIDES, S., KALININ, S. V. and BORISEVICH, A. Y. (2014). Direct observa-

- tion of ferroelectric field effect and vacancy-controlled screening at the $\text{BiFeO}_3/\text{La}_x\text{Sr}_{1-x}\text{MnO}_3$ interface. *Nat. Mater.* **13** 1019–1025.
- LIU, J. and YE, J. (2010). Moreau–Yosida regularization for grouped tree structure learning. In *Advances in Neural Information Processing Systems* 1459–1467.
- MODY, C. C. M. (2011). *Instrumental Community: Probe Microscopy and the Path to Nanotechnology*. MIT Press, Cambridge, MA.
- NELLIST, P. and PENNYCOOK, S. (2000). The principles and interpretations of annular dark-field Z-contrast imaging. *Adv. Imaging Electron Phys.* **113** 148–204.
- NELSON, C. T., WINCHESTER, B., ZHANG, Y., KIM, S.-J., MELVILLE, A., ADAMO, C., FOLKMAN, C. M., BAEK, S.-H., EOM, C.-B., SCHLOM, D. G. et al. (2011). Spontaneous vortex nanodomain arrays at ferroelectric heterointerfaces. *Nano Lett.* **11** 828–834.
- REZATOFIHI, S. H., HARTLEY, R. and HUGHES, W. E. (2012). A new approach for spot detection in total internal reflection fluorescence microscopy. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)* 860–863.
- SAGE, D., NEUMANN, F. R., HEDIGER, F., GASSER, S. M. and UNSER, M. (2005). Automatic tracking of individual fluorescence particles: Application to the study of chromosome dynamics. *IEEE Trans. Image Process.* **14** 1372–1383.
- SANG, X., LUPINI, A. R., UNOCIC, R. R., CHI, M., BORISEVICH, A. Y., KALININ, S. V., ENDEVE, E., ARCHIBALD, R. K. and JESSE, S. (2016a). Dynamic scan control in STEM: Spiral scans. *Adv. Struct. Chem. Imaging* **2** 6.
- SANG, X., XIE, Y., LIN, M.-W., ALHABEB, M., VAN AKEN, K. L., GOGOTSI, Y., KENT, P. R., XIAO, K. and UNOCIC, R. R. (2016b). Atomic defects in monolayer titanium carbide ($\text{Ti}_3\text{C}_2\text{T}_x$) MXene. *ACS Nano* **10** 9193–9200.
- SHE, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electron. J. Stat.* **3** 384–415.
- SHE, Y. (2010). Sparse regression with exact clustering. *Electron. J. Stat.* **4** 1055–1096. [MR2727453](#)
- SHE, Y., WANG, J., LI, H. and WU, D. (2013). Group iterative spectrum thresholding for super-resolution sparse spectral selection. *IEEE Trans. Signal Process.* **61** 6371–6386. [MR3148325](#)
- SHIJU, N. R. and GULIANTS, V. V. (2009). Recent developments in catalysis using nanostructured materials. *Appl. Catal., A Gen.* **356** 1–17.
- SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245.
- SLAWSKI, M. and HEIN, M. (2013). Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electron. J. Stat.* **7** 3004–3056.
- SMAL, I., NIESSEN, W. and MEIJERING, E. (2008). A new detection scheme for multiple object tracking in fluorescence microscopy by joint probabilistic data association filtering. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 264–267.
- VASUDEVAN, R. K., BELIANINOV, A., GIANFRANCESCO, A. G., BADDORF, A. P., TSELEV, A., KALININ, S. V. and JESSE, S. (2015). Big data in reciprocal space: Sliding fast Fourier transforms for determining periodicity. *Appl. Phys. Lett.* **106** 091601.
- VINCENT, L. (1993). Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms. *IEEE Trans. Image Process.* **2** 176–201.
- YANKOVICH, A. B., BERKELS, B., DAHMEN, W., BINEV, P., SANCHEZ, S. I., BRADLEY, S. A., LI, A., SZLUFARSKA, I. and VOYLES, P. M. (2014). Picometre-precision analysis of scanning transmission electron microscopy images of platinum nanocatalysts. *Nat. Commun.* **5** 4155.
- YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. [MR2212574](#)

A SPATIO-TEMPORAL MODELING FRAMEWORK FOR WEATHER RADAR IMAGE DATA IN TROPICAL SOUTHEAST ASIA

BY XIAO LIU, VIKNESWARAN GOPAL AND JAYANT KALAGNANAM

University of Arkansas, National University of Singapore, and IBM Thomas J. Watson Research Center

Tropical storms are known to be highly chaotic and extremely difficult to predict. In tropical countries such as Singapore, the official lead time for the warnings of heavy storms is usually between 15 and 45 minutes because weather systems develop quickly and are of very short lifespan. A single thunderstorm cell, for example, typically lives for less than an hour. Weather radar echoes, correlated in both space and time, provide a rich source of information for short-term precipitation nowcasting. Based on a large dataset of 276 tropical storms events, this paper investigates a spatio-temporal modeling approach for two-dimensional radar reflectivity (echo) fields. Under a Lagrangian integration scheme, we model the radar reflectivity field by a spatio-temporal conditional autoregressive process with two components. The first component is the dynamic velocity field which determines the motion of the storm, and the second component governs the growth or decay of the returned radar echoes. The proposed method is demonstrated and compared with existing methods using real radar image data collected from a number of 276 tropical storm events from 2010 to 2011 in Singapore. The numerical comparison results show the advantage of the proposed method, in terms of the mean-squared-error, in modeling small-scale localized convective weather systems based on the 77 inter-monsoon season thunderstorm events.

REFERENCES

- ASSUNÇÃO, R. and KRAINSKI, E. (2009). Neighborhood dependence in Bayesian spatial models. *Biom. J.* **51** 851–869. [MR2751717](#)
- 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](#)
- BERTSEKAS, D. P. (1982). *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, Boston. [MR0690767](#)
- BESAG, J. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746. [MR1380811](#)
- BOWLER, N. E. H., PIERCE, C. E. and SEED, A. (2004). Development of a precipitation nowcasting algorithm based upon optical flow techniques. *J. Hydrol.* **288** 74–91.
- BOWLER, N. E., PIERCE, C. E. and SEED, A. W. (2006). A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Q. J. R. Meteorol. Soc.* **132** 2127–2155.

Key words and phrases. Spatial-temporal modeling, precipitation forecast, space-time conditional autoregressive model, radar image analysis.

- BROWN, P. E., DIGGLE, P. J., LORD, M. E. and YOUNG, P. C. (2001). Space-time calibration of radar rainfall data. *J. Roy. Statist. Soc. Ser. C* **50** 221–241. [MR1833274](#)
- BROWNING, K. A., COLLIER, C. G., LARK, P. R., MENMUIR, P., MONK, G. A. and OWENS, R. G. (1982). On the forecasting of frontal rain using a weather radar network. *Mon. Weather Rev.* **110** 534–552.
- CARLIN, B. P. and BANERJEE, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data (with discussion). In *Bayesian Statistics, 7 (Tenerife, 2002)* 45–63. Oxford Univ. Press, New York. [MR2003166](#)
- CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York. [MR1239641](#)
- DIXON, M. and WIENER, G. (1993). TITAN: Thunderstorm identification, tracking, analysis, and nowcasting—a radar-based methodology. *J. Atmos. Ocean. Technol.* **10** 785–797.
- FUENTES, M., REICH, B. and LEE, G. (2008). Spatial-temporal mesoscale modeling of rainfall intensity using gage and radar data. *Ann. Appl. Stat.* **2** 1148–1169. [MR2655653](#)
- GELPKE, V. and KÜNSCH, H. R. (2001). Estimation of motion from sequences of images. In *Spatial Statistics: Methodological Aspects and Applications. Lect. Notes Stat.* **159** 141–167. Springer, New York. [MR1847746](#)
- GERMANN, U. and ZAWADZKI, I. (2002). Scale-dependence of the predicability of precipitation from continental radar images. Part I: Description of the methodology. *Mon. Weather Rev.* **130** 2859–2873.
- HAINING, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge Univ. Press, Cambridge.
- HAN, L., FU, S. X., ZHAO, L. F., ZHENG, Y. G., WANG, H. Q. and LIN, Y. J. (2009). 3D convective storm identification, tracking, and forecasting—an enhanced TITAN algorithm. *J. Atmos. Ocean. Technol.* **26** 719–732.
- HARTIGAN, J. A. and WONG, M. A. (1979). A K-means clustering algorithm. *Appl. Stat.* **28** 100–108.
- HORN, B. K. P. and SCHUNCK, B. G. (1981). Determining optical flow. *Artificial Intelligence* **17** 185–203.
- JIN, X., CARLIN, B. P. and BANERJEE, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics* **61** 950–961. [MR2216188](#)
- LEESE, J. A., NOVAK, C. S. and CLARK, B. B. (1971). An automated technique for obtaining cloud motion from geosynchronous satellite data using cross correlation. *J. Appl. Meteorol.* **10** 118–132.
- LI, P. and LAI, S. T. (2004). Short-range quantitative precipitation forecasting in Hong Kong. *J. Hydrol.* **288** 189–209.
- LI, L., SCHMID, W. and JOSS, J. (1995). Nowcasting of motion and growth of precipitation with radar over a complex orography. *J. Appl. Meteorol.* **34** 1286–1300.
- MARIELLA, L. and TARANTINO, M. (2010). Spatial temporal conditional auto-regressive model: A new autoregressive matrix. *Aust. J. Stat.* **39** 223–244.
- MARSHALL, J. S. and PALMER, W. M. (1948). The distribution of raindrops with size. *J. Meteorol.* **5** 165–166.
- NATIONAL ENVIRONMENTAL AGENCY SINGAPORE (2014). Weather Statistics. Available at <http://app2.nea.gov.sg/weather-climate/climate-information/weather-statistics>.
- NATIONAL ENVIRONMENTAL AGENCY SINGAPORE (2017). Challenges in Weather Forecasting. Available at <http://www.nea.gov.sg/training-knowledge/weather-climate/weather-forecast/challenges-in-weather-forecasting/>.
- RADHAKRISHNA, B., ZAWADZKI, I. and FABRY, F. (2012). Predictability of precipitation from continental radar images. Part V: Growth and decay. *J. Atmos. Sci.* **69** 3336–3349.
- RINEHART, R. E. and GARVEY, E. T. (1978). Three-dimensional storm motion detection by conventional weather radar. *Nature* **273** 287–289.
- ROYAL METEOROLOGICAL INSTITUTE OF BELGIUM (2008). Quantitative Precipitation Forecasts based on radar observations: Principles, algorithms and operational systems, 2008/0224/52.

- SEED, A. W. (2003). A dynamic and spatial scaling approach to advection forecasting. *J. Appl. Meteorol.* **42** 381–388.
- SEED, A. W., PIERCE, C. E. and NORMAN, K. (2013). Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme. *Water Resour. Res.* **49** 6624–6641.
- SIGRIST, F., KÜNSCH, H. R. and STAHEL, W. A. (2012). A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. *Ann. Appl. Stat.* **6** 1452–1477. [MR3058671](#)
- STANIFORTH, A. and COTE, J. (1991). Semi-Lagrangian integration schemes for atmospheric models—a review. *Mon. Weather Rev.* **119** 2206–2223.
- STERN, H. and CRESSIE, N. (2000). Posterior predictive model checks for disease mapping models. *Stat. Med.* **19** 2377–2397.
- STROUD, J. R., MÜLLER, P. and SANSÓ, B. (2001). Dynamic models for spatiotemporal data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 673–689. [MR1872059](#)
- TESTIK, F. and GEBREMICHAEL, M. (2013). *Rainfall: State of the Science*. American Geophysical, Union.
- WALL, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *J. Statist. Plann. Inference* **121** 311–324. [MR2038824](#)
- WILSON, J. W., CROOK, N. A., MUELLER, C. K., SUN, J. and DIXON, M. (1998). Nowcasting thunderstorms: A status report. *Bull. Am. Meteorol. Soc.* **78** 2079–2099.
- WOLFSON, M. M., FORMAN, B. E., HALLOWELL, R. G. and MOORE, M. P. (1999). *The Growth and Decay Storm Tracker*, The 8th Conference on Aviation, Range, and Aerospace Meteorology.
- XU, K., WIKLE, C. K. and FOX, N. I. (2005). A kernel-based spatio-temporal dynamical model for nowcasting weather radar reflectivities. *J. Amer. Statist. Assoc.* **100** 1133–1144. [MR2236929](#)

MULTIVARIATE INTEGER-VALUED TIME SERIES WITH FLEXIBLE AUTOCOVARIANCES AND THEIR APPLICATION TO MAJOR HURRICANE COUNTS¹

BY JAMES LIVSEY^{*}, ROBERT LUND^{†,2}, STEFANOS KECHAGIAS[‡]
 AND VLADAS PIPIRAS^{§,3}

United States Census Bureau^{}, Clemson University[†], SAS Institute[‡]
 and University of North Carolina, Chapel Hill[§]*

This paper examines a bivariate count time series with some curious statistical features: Saffir–Simpson Category 3 and stronger annual hurricane counts in the North Atlantic and eastern Pacific Ocean Basins. As land and ocean temperatures on our planet warm, an intense climatological debate has arisen over whether hurricanes are becoming more numerous, or whether the strengths of the individual storms are increasing. Recent literature concludes that an increase in hurricane counts occurred in the Atlantic Basin circa 1994. This increase persisted through 2012; moreover, the 1994–2012 period was one of relative inactivity in the eastern Pacific Basin. When Atlantic activity eased in 2013, heavy activity in the eastern Pacific Basin commenced. When examined statistically, a Poisson white noise model for the annual severe hurricane counts is difficult to resoundingly reject. Yet, decadal cycles (longer term dependence) in the hurricane counts is plausible. This paper takes a statistical look at the issue, developing a stationary multivariate count time series model with Poisson marginal distributions and a flexible autocovariance structure. Our auto- and cross-correlations can be negative and have long-range dependence; features that most previous count models cannot achieve in tandem. Our model is new in the literature and is based on categorizing and super-positioning multivariate Gaussian time series. We derive the autocovariance function of the model and propose a method to estimate model parameters. In the end, we conclude that severe hurricane counts are indeed negatively correlated across the two ocean basins. Some evidence for long-range dependence is also presented; however, with only a 49-year record, this issue cannot be definitively judged without additional data.

REFERENCES

- ALZAID, A. A. and AL-OSH, M. (1990). An integer-valued p th-order autoregressive structure (INAR(p)) process. *J. Appl. Probab.* **27** 314–324. [MR1052303](#)
- BARNDORFF-NIELSEN, O. E., LUNDE, A., SHEPHARD, N. and VERAART, A. E. D. (2014). Integer-valued trawl processes: A class of stationary infinitely divisible processes. *Scand. J. Stat.* **41** 693–724. [MR3249424](#)
- BERAN, J., FENG, Y., GHOSH, S. and KULIK, R. (2013). *Long-Memory Processes*. Springer, Heidelberg.

Key words and phrases. Count time series, hurricanes, long-range dependence, multivariate, negative autocorrelation, Poisson.

- BLIGHT, P. A. (1989). Time series formed from the superposition of discrete renewal processes. *J. Appl. Probab.* **26** 189–195. [MR0981265](#)
- BROCKWELL, P. J. and DAVIS, R. A. (2006). *Time Series: Theory and Methods*. Springer, New York. Reprint of the second (1991) edition. [MR2839251](#)
- CHU, P. S. and ZHAO, Z. (2004). Bayesian change-point analysis of tropical cyclone activity: The Central North Pacific case. *J. Climate* **17** 4893–4901.
- CUI, Y. and LUND, R. (2009). A new look at time series of counts. *Biometrika* **96** 781–792.
- CZADO, C., GNEITING, T. and HELD, L. (2009). Predictive model assessment for count data. *Biometrics* **65** 1254–1261.
- DAVIS, R. A. and DUNSMUIR, W. T. M. (2016). State space models for count time series. In *Handbook of Discrete-Valued Time Series* (R. A. Davis, S. Holan, R. B. Lund and N. Ravishanker, eds.) CRC Press, Boca Raton, FL.
- DAVIS, R. A., HOLAN, S. H., LUND, R. B. and RAVISHANKER, N., eds. (2016). *Handbook of Discrete-Valued Time Series*. CRC Press, Boca Raton.
- DOUKHAN, P., OPPENHEIM, G. and TAQQU, M. S. (2003). *Theory and Applications of Long-Range Dependence*. Birkhäuser, Boston, MA.
- DUNSMUIR, W. T. M. (2016). Generalized linear autoregressive moving average models. In *Handbook of Discrete-Valued Time Series* (R. A. Davis, S. Holan, R. B. Lund and R. N. Ravishanker, eds.) CRC Press, Boca Raton, FL.
- ELSNER, J. and JAGGER, T. H. (2006). Prediction models for annual US hurricane counts. *J. Climate* **19** 2935–2952.
- ELSNER, J., JAGGER, T. and NIU, X. F. (2000). Changes in the rates of North Atlantic major hurricane activity during the 20th century. *Geophys. Res. Lett.* **27** 1743–1746.
- ELSNER, J. B. and KOCHER, B. (2000). Global tropical cyclone activity: A link to the North Atlantic Oscillation. *Geophys. Res. Lett.* **27** 129–132.
- ELSNER, J., KOSSIN, J. P. and JAGGER, T. H. (2008). The increasing intensity of the strongest tropical cyclones. *Nature* **455** 92–95.
- ENCISO-MORA, V., NEAL, P. and SUBBA RAO, T. (2009). Efficient order selection for integer-valued ARMA processes. *J. Time Series Anal.* **30** 1–18.
- FOKIANOS, K. and KEDEM, B. (2003). Regression theory for categorical time series. *Statist. Sci.* **18** 357–376. [MR2061915](#)
- GIRAITIS, L., KOUL, H. L. and SURGAILIS, D. (2012). *Large Sample Inference for Long Memory Processes*. Imperial College Press, London.
- GOLDENBERG, S. B., LANDSEA, C. W. and MESTAS-NUNEZ, A. M. (2001). The recent increase in Atlantic hurricane activity: Causes and implications. *Science* **293** 474–479.
- GRAY, W. M. (1984). Atlantic seasonal hurricane frequency. Part I: El Nino and 30 mb quasi-biennial oscillation influences. *Mon. Weather Rev.* **112** 1649–1668.
- HASSELMANN, K. (1976). Stochastic climate models part I. Theory. *Tellus* **28** 473–485.
- HELGASON, H., PIPIRAS, V. and ABRY, P. (2011). Fast and exact synthesis of stationary multivariate Gaussian time series using circulant embedding. *Signal Process.* **91** 1123–1133.
- JIA, Y. and LUND, R. B. (2016). Superpositioned stationary count time series. *J. Appl. Probab.*. To appear.
- JOE, H. (1996). Time series models with univariate margins in the convolution-closed infinitely divisible class. *J. Appl. Probab.* **33** 664–677. [MR1401464](#)
- KACHOUR, M. and YAO, J. F. (2009). First-order rounded integer-valued autoregressive (RINAR(1)) process. *J. Time Series Anal.* **30** 417–448. [MR2536061](#)
- KARLIS, D. (2016). Models for multivariate count time series. In *Handbook of Discrete-Valued Time Series* (R. A. Davis, S. Holan, R. B. Lund and N. Ravishanker, eds.) 407–424. CRC Press, Boca Raton, FL.
- KARLIS, D. and MELIGKOTSIDOU, L. (2007). Finite mixtures of multivariate Poisson distributions with application. *J. Statist. Plann. Inference* **137** 1942–1960.

- KECHAGIAS, S. and PIPIRAS, V. (2015). Definitions and representations of multivariate long-range dependent time series. *J. Time Series Anal.* **36** 1–25. [MR3300202](#)
- KECHAGIAS, S. and PIPIRAS, V. (2017). Identification, estimation and applications of a bivariate long-range dependent time series model with general phase. Preprint.
- KERSS, A., LEONENKO, N. and SIKORSKII, A. (2014). Fractional Skellam processes with applications to finance. *Fract. Calc. Appl. Anal.* **17** 532–551.
- KOLASSA, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *Int. J. Forecast.* **32** 788–803.
- LUND, R. B., HOLAN, S. H. and LIVSEY, J. (2016). Long memory discrete-valued time series. In *Handbook of Discrete-Valued Time Series* (R. A. Davis, S. Holan, R. B. Lund and N. Ravishanker, eds.) 447–458. CRC Press, Boca Raton, FL.
- LUND, R. B. and LIVSEY, J. (2016). Renewal-based count time series. In *Handbook of Discrete-Valued Time Series* (R. A. Davis, S. Holan, R. B. Lund and N. Ravishanker, eds.) 101. CRC Press, Boca Raton.
- LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin. [MR2172368](#)
- MACDONALD, I. L. and ZUCCHINI, W. (2016). Hidden Markov models for discrete-valued time series. In *Handbook of Discrete-Valued Time Series* (R. Davis, S. Holan, R. B. Lund and N. Ravishanker, eds.) CRC Press, Boca Raton, FL.
- MCDONNELL, K. A. and HOLBROOK, N. J. (2004). A Poisson regression model of tropical cyclogenesis for the Australian-southwest Pacific Ocean region. *Weather Forecast.* **19** 440–455.
- MCKENZIE, E. (2003). Discrete variate time series. In *Stochastic Processes: Modelling and Simulation. Handbook of Statistics* **21** 573–606. North-Holland, Amsterdam.
- MOOLEY, D. A. (1981). Applicability of the Poisson probability model to severe cyclonic storms striking the coast around the bay of bengal. *Sankhya* **43** 187–197.
- MOONEY, C. C. (2007). *Storm World. Hurricanes, Politics, and the Battle over Global Warming*. Harcourt, New York.
- MUDELSEE, M. (2013). *Climate Time Series Analysis*. Springer, Berlin.
- NEAL, P. and SUBBA RAO, T. (2007). MCMC for integer-valued ARMA processes. *J. Time Series Anal.* **28** 92–110. [MR2332852](#)
- PALMA, W. (2007). *Long-Memory Time Series*. Wiley, New Jersey.
- PARISI, F. and LUND, R. B. (2000). Seasonality and return periods of landfalling Atlantic basin hurricanes. *Aust. N. Z. J. Stat.* **42** 271–282.
- PARK, K. and WILLINGER, W. (2000). *Self-Similar Network Traffic and Performance Evaluation*. Wiley Online Library.
- PERCIVAL, D. B., OVERLAND, J. E. and MOFJELD, H. O. (2001). Interpretation of North Pacific variability as a short-and long-memory process. *J. Climate* **14** 4545–4559.
- PIPIRAS, V. and TAQQU, M. S. (2017). *Long-Range Dependence and Self-Similarity* **45**. Cambridge Univ. Press, Cambridge.
- QUORESHI, A. M. M. S. (2014). A long-memory integer-valued time series model, INARFIMA, for financial application. *Quant. Finance* **14** 2225–2235.
- ROBBINS, M. W., LUND, R. B., GALLAGHER, C. M. and LU, Q. (2011). Changepoints in the North Atlantic tropical cyclone record. *J. Amer. Statist. Assoc.* **106** 89–99. [MR2816704](#)
- ROBINSON, P. M. (2003). Long-memory time series. In *Time Series with Long Memory. Adv. Texts Econometrics* 4–32. Oxford Univ. Press, Oxford. [MR2077520](#)
- SELA, R. J. (2010). Three essays in econometrics: Multivariate long memory time series and applying regression trees to longitudinal data. Ph.D. thesis, New York Univ.
- SKELLAM, J. G. (1946). The frequency distribution of the difference between two Poisson variates belonging to different populations. *J. R. Stat. Soc.* **109** 296.
- SOLOW, A. (1989). Statistical modeling of storm counts. *J. Climate* **2** 131–136.

- STEUTEL, F. W. and VAN HARN, K. (1979). Discrete analogues of self-decomposability and stability. *Ann. Probab.* **7** 893–899.
- THOMPSON, M. L. and GUTTORP, P. (1986). A probability model for severe cyclonic storms striking the coast around the bay of bengal. *Mon. Weather Rev.* **114** 2267–2271.
- VAN VLECK, J. H. and MIDDLETON, D. (1966). The spectrum of clipped noise. *Proc. IEEE* **54** 2–19.
- VAROTSOS, C. and EFSTATHIOU, M. N. (2013). Is there any long-term memory effect in the tropical cyclones? *Theor. Appl. Climatol.* **114** 643–650.
- VILLARINI, G., VECCHI, G. A. and SMITH, J. A. (2010). Modeling the dependence of tropical storm counts in the North Atlantic basin on climate indices. *Mon. Weather Rev.* **138** 2681–2705.
- XIAO, S., KOTTAS, A. and SANSÓ, B. (2015). Modeling for seasonal marked point processes: An analysis of evolving hurricane occurrences. *Ann. Appl. Stat.* **9** 353–382. [MR3341119](#)
- YUAN, N., FU, Z. and LIU, S. (2014). Extracting climate memory using fractional integrated statistical model: A new perspective on climate prediction. *Sci. Rep.* **4**.

STOCHASTIC SIMULATION OF PREDICTIVE SPACE-TIME SCENARIOS OF WIND SPEED USING OBSERVATIONS AND PHYSICAL MODEL OUTPUTS¹

BY JULIE BESSAC^{*}, EMIL CONSTANTINESCU^{*,†} AND MIHAI ANITESCU^{*,†}

Argonne National Laboratory^{} and the University of Chicago[†]*

We propose a statistical space–time model for predicting atmospheric wind speed based on deterministic numerical weather predictions and historical measurements. We consider a Gaussian multivariate space–time framework that combines multiple sources of past physical model outputs and measurements in order to produce a probabilistic wind speed forecast within the prediction window. We illustrate this strategy on wind speed forecasts during several months in 2012 for a region near the Great Lakes in the United States. The results show that the prediction is improved in the mean-squared sense relative to the numerical forecasts as well as in probabilistic scores. Moreover, the samples are shown to produce realistic wind scenarios based on sample spectra and space–time correlation structure.

REFERENCES

- AILLIOT, P., FRÉNOT, E. and MONBET, V. (2006). Long term object drift forecast in the ocean with tide and wind. *Multiscale Model. Simul.* **5** 514–531. [MR2247761](#)
- ANDERSON, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate* **9** 1518–1530.
- ANITESCU, M., CHEN, J. and WANG, L. (2012). A matrix-free approach for solving the parametric Gaussian process maximum likelihood problem. *SIAM J. Sci. Comput.* **34** A240–A262.
- APANASOVICH, T. and GENTON, M. G. (2010). Cross-covariance functions for multivariate random fields based on latent dimensions. *Biometrika* **97** 15–30. [MR2594414](#)
- BAO, L., GNEITING, T., GRIMIT, E. P., GUTTORP, P. and RAFTERY, A. E. (2010). Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Mon. Weather Rev.* **138** 1811–1821.
- BARAN, S. (2014). Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Comput. Statist. Data Anal.* **75** 227–238.
- BARAN, S. and LERCH, S. (2015). Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind-speed forecasting. *Q. J. R. Meteorol. Soc.* **141** 2289–2299.
- BARAN, S. and LERCH, S. (2016). Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics* **27** 116–130. [MR3481324](#)
- BERLINER, M. (2000). Hierarchical Bayesian modeling in the environmental sciences. *ASta Adv. Stat. Anal.* **84** 141–153.
- 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.

Key words and phrases. Hierarchical Gaussian model, multiple data sources, predictive scenarios, spatio-temporal wind speed.

- BOUALLEGUE, Z. B., HEPPELMANN, T., THEIS, S. E. and PINSON, P. (2016). Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach. *Mon. Weather Rev.* **144** 4737–4750.
- BOUROTTE, M., ALLARD, D. and PORCU, E. (2016). A flexible class of non-separable cross-covariance functions for multivariate space–time data. *Spat. Stat.* **18** 125–146. [MR3573273](#)
- BRISSON, N., GARY, C., JUSTES, E., ROCHE, R., MARY, B., RIPOCHE, D., ZIMMER, D., SIERRA, J., BERTUZZI, P. and BURGER, P. (2003). An overview of the crop model STICS. *Eur. J. Agron.* **18** 309–332.
- BROWN, B. G., KATZ, R. W. and MURPHY, A. H. (1984). Time series models to simulate and forecast wind speed and wind power. *J. Clim. Appl. Meteorol.* **23** 1184–1195.
- 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.
- CONSTANTINESCU, E. M. and ANITESCU, M. (2013). Physics-based covariance models for Gaussian processes with multiple outputs. *Int. J. Uncertain. Quantif.* **3** 47–71. [MR3044910](#)
- CONSTANTINESCU, E., ZAVALA, V., ROCKLIN, M., LEE, S. and ANITESCU, M. (2011). A computational framework for uncertainty quantification and stochastic optimization in unit commitment with wind power generation. *IEEE Trans. Power Syst.* **26** 431–441.
- COWLES, M. K., ZIMMERMAN, D. L., CHRIST, A. and MCGINNIS, D. L. (2002). Combining snow water equivalent data from multiple sources to estimate spatio-temporal trends and compare measurement systems. *J. Agric. Biol. Environ. Stat.* **7** 536–557.
- CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ. [MR2848400](#)
- DAWID, A. P. and SEBASTIANI, P. (1999). Coherent dispersion criteria for optimal experimental design. *Ann. Statist.* **27** 65–81.
- FANSHAWE, T. R. and DIGGLE, P. J. (2012). Bivariate geostatistical modelling: A review and an application to spatial variation in Radon concentrations. *Environ. Ecol. Stat.* **19** 139–160. [MR2946168](#)
- FELDMANN, K., SCHEUERER, M. and THORARINSDOTTIR, T. L. (2015). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Monthly 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.
- FUENTES, M., CHEN, L., DAVIS, J. M. and LACKMANN, G. M. (2005). Modeling and predicting complex space–time structures and patterns of coastal wind fields. *Environmetrics* **16** 449–464.
- 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.
- GENTON, M. G. and KLEIBER, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statist. Sci.* **30** 147–163. [MR3353096](#)
- GLAHN, H. R. and LOWRY, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.* **11** 1203–1211.
- GNEITING, T., RAFTERY, A. E., WESTVELD, A. H. III and GOLDMAN, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133** 1098–1118.
- GNEITING, T., LARSON, K., WESTRICK, K., GENTON, M. G. and ALDRICH, E. (2006). Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space–time method. *J. Amer. Statist. Assoc.* **101** 968–979.
- GNEITING, T., STANBERRY, L. I., GRIMIT, E. P., HELD, L. and JOHNSON, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST* **17** 211–235.

- HAMILL, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Weather Rev.* **129** 550–560.
- HERING, A. S. and GENTON, M. G. (2010). Powering up with space–time wind forecasting. *J. Amer. Statist. Assoc.* **105** 92–104.
- KANG, E. L., CRESSIE, N. and SAIN, S. R. (2012). Combining outputs from the North American regional climate change assessment program by using a Bayesian hierarchical model. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 291–313.
- KAZOR, K. and HERING, A. S. (2015). The role of regimes in short-term wind speed forecasting at multiple wind farms. *Stat* **4** 271–290. [MR3429323](#)
- LERCH, S. and THORARINSDOTTIR, T. L. (2013). Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A* **65** 21206.
- LI, N., UCKUN, C., CONSTANTINESCU, E., BIRGE, J., HEDMAN, K. and BOTTERUD, A. (2015). Flexible operation of batteries in power system scheduling with renewable energy. *IEEE Trans. Sustain. Energy* **7** 685–696.
- PALMER, T. (2014). More reliable forecasts with less precise computations: A fast-track route to cloud-resolved weather and climate simulators? *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **372** 20130391.
- PAPAVASILIOU, A., OREN, S. S. and ROUNTREE, B. (2015). Applying high performance computing to transmission-constrained stochastic unit commitment for renewable energy integration. *IEEE Trans. Power Syst.* **30** 1109–1120.
- PINSON, P. (2013). Wind energy: Forecasting challenges for its operational management. *Statist. Sci.* **28** 564–585.
- PINSON, P. and GIRARD, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Appl. Energy* **96** 12–20.
- PINSON, P. and TASTU, J. (2013). Discrimination ability of the energy score. Technical University of Denmark. Tech. rep.
- PINSON, P., CHRISTENSEN, L. E. A., MADSEN, H., SORENSEN, P. E., DONOVAN, M. H. and JENSEN, L. E. (2008). Regime-switching modelling of the fluctuations of offshore wind generation. *J. Wind Eng. Ind. Aerodyn.* **96** 2327–2347.
- PINSON, P., MADSEN, H., NIELSEN, H., PAPAETHYMIU, G. and KLÖCKL, B. (2009). From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy* **12** 51–62.
- RAFTERY, A. E., GNEITING, T., BALABDAOUI, F. and POLAKOWSKI, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133** 1155–1174.
- ROYLE, J. A. and BERLINER, L. M. (1999). A hierarchical approach to multivariate spatial modeling and prediction. *J. Agric. Biol. Environ. Stat.* **4** 29–56. [MR1812239](#)
- ROYLE, J., BERLINER, L., WIKLE, C. and MILLIFF, R. (1999). A hierarchical spatial model for constructing wind fields from scatterometer data in the Labrador Sea. In *Case Studies in Bayesian Statistics* 367–382. Springer, Berlin.
- SCHEFZIK, R., THORARINSDOTTIR, T. L. and GNEITING, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statist. Sci.* **28** 616–640.
- SCHEUERER, M. and HAMILL, T. M. (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Mon. Weather Rev.* **143** 1321–1334.
- SCHEUERER, M. and MÖLLER, D. (2015). Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Ann. Appl. Stat.* **9** 1328–1349. [MR3418725](#)
- SCHUHEN, N., THORARINSDOTTIR, T. L. and GNEITING, T. (2012). Ensemble model output statistics for wind vectors. *Mon. Weather Rev.* **140** 3204–3219.
- SHUMWAY, R. and STOFFER, D. (2010). *Time Series Analysis and Its Applications: With R Examples*. Springer Science & Business Media.

- SKAMAROCK, W., KLEMP, J., DUDHIA, J., GILL, D., BARKER, D., DUDA, M., HUANG, X.-Y., WANG, W. and POWERS, J. (2008). A description of the Advanced Research WRF Version 3 Tech. Rep. Tech Notes-475+ STR, NCAR.
- SLOUGHTER, J. M. L., GNEITING, T. and RAFTERY, A. E. (2010). Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Amer. Statist. Assoc.* **105** 25–35.
- SLOUGHTER, J. M. L., GNEITING, T. and RAFTERY, A. E. (2013). Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Mon. Weather Rev.* **141** 2107–2119.
- SMITH, L. A. and HANSEN, J. A. (2004). Extending the limits of ensemble forecast verification with the minimum spanning tree. *Mon. Weather Rev.* **132** 1522–1528.
- STEIN, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York. [MR1697409](#)
- STEIN, M., CHEN, J. and ANITESCU, M. (2012). Difference filter preconditioning for large covariance matrices. *SIAM J. Matrix Anal. Appl.* **33** 52–72.
- THORARINSDOTTIR, T. L. and GNEITING, T. (2010). Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Statist. Soc. Ser. A* **173** 371–388.
- THORARINSDOTTIR, T. L. and JOHNSON, M. S. (2012). Probabilistic wind gust forecasting using non-homogeneous Gaussian regression. *Mon. Weather Rev.* **140** 889–897.
- THORARINSDOTTIR, T. L., SCHEUERER, M. and HEINZ, C. (2016). Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *J. Comput. Graph. Statist.* **25** 105–122.
- WILKS, D. S. (2015). Multivariate ensemble Model Output Statistics using empirical copulas. *Q. J. R. Meteorol. Soc.* **141** 945–952.

A MULTI-RESOLUTION MODEL FOR NON-GAUSSIAN RANDOM FIELDS ON A SPHERE WITH APPLICATION TO IONOSPHERIC ELECTROSTATIC POTENTIALS

BY MINJIE FAN^{*,1}, DEBASHIS PAUL^{*,2}, THOMAS C. M. LEE^{*,3} AND TOMOKO MATSUO^{†,1}

University of California, Davis^{} and University of Colorado, Boulder[†]*

Gaussian random fields have been one of the most popular tools for analyzing spatial data. However, many geophysical and environmental processes often display non-Gaussian characteristics. In this paper, we propose a new class of spatial models for non-Gaussian random fields on a sphere based on a multi-resolution analysis. Using a special wavelet frame, named *spherical needlets*, as building blocks, the proposed model is constructed in the form of a sparse random effects model. The spatial localization of needlets, together with carefully chosen random coefficients, ensure the model to be non-Gaussian and isotropic. The model can also be expanded to include a spatially varying variance profile. The special formulation of the model enables us to develop efficient estimation and prediction procedures, in which an adaptive MCMC algorithm is used. We investigate the accuracy of parameter estimation of the proposed model, and compare its predictive performance with that of two Gaussian models by extensive numerical experiments. Practical utility of the proposed model is demonstrated through an application of the methodology to a data set of high-latitude ionospheric electrostatic potentials, generated from the LFM-MIX model of the magnetosphere-ionosphere system.

REFERENCES

- ANDREWS, D. F. and MALLOWS, C. L. (1974). Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B* **36** 99–102. [MR0359122](#)
- ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* **18** 343–373.
- ATKINSON, K. and HAN, W. (2012). *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction. Lecture Notes in Math.* **2044**. Springer, Heidelberg. [MR2934227](#)
- BALDI, P., KERKYCHARIAN, G., MARINUCCI, D. and PICARD, D. (2009). Asymptotics for spherical needlets. *Ann. Statist.* **37** 1150–1171.
- BARNDORFF-NIELSEN, O. (1979). Models for non-Gaussian variation, with applications to turbulence. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **368** 501–520. [MR0554672](#)
- BERG, J., NATARAJAN, A., MANN, J. and PATTON, E. G. (2016). Gaussian vs non-Gaussian turbulence: Impact on wind turbine loads. *Wind Energy* **19** 1975–1989.
- BHATTACHARYA, A., CHAKRABORTY, A. and MALLICK, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika* **103** 985–991. [MR3620452](#)
- CHU, J.-H., CLYDE, M. A. and LIANG, F. (2009). Bayesian function estimation using continuous wavelet dictionaries. *Statist. Sinica* **19** 1419–1438.

Key words and phrases. Non-Gaussian random field, multi-resolution analysis, isotropic process on a sphere, MCMC, ionospheric electrostatic potential, LFM-MIX model.

- CODRESCU, M. V., FULLER-ROWELL, T. J. and FOSTER, J. C. (1995). On the importance of E-field variability for Joule heating in the high-latitude thermosphere. *Geophys. Res. Lett.* **22** 2393–2396.
- CODRESCU, M. V., FULLER-ROWELL, T. J., FOSTER, J. C., HOLT, J. M. and CARIGLIA, S. J. (2000). Electric field variability associated with the Millstone Hill electric field model. *J. Geophys. Res.* **105** 5265–5273.
- COUSINS, E. D. P., MATSUO, T. and RICHMOND, A. D. (2013a). SuperDARN assimilative mapping. *J. Geophys. Res.* **118** 7954–7962.
- COUSINS, E. D. P., MATSUO, T. and RICHMOND, A. D. (2013b). Mesoscale and large-scale variability in high-latitude ionospheric convection: Dominant modes and spatial/temporal coherence. *J. Geophys. Res.* **118** 7895–7904.
- COUSINS, E. D. P. and SHEPHERD, S. G. (2012). Statistical characteristics of small-scale spatial and temporal electric field variability in the high-latitude ionosphere. *J. Geophys. Res.* **117** A03317.
- CRESSIE, N. (1993). *Statistics for Spatial Data*. Wiley, New York, NY.
- CRESSIE, N. and JOHANNESSON, G. (2008). Fixed rank kriging for very large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 209–226.
- CRESSIE, N., SHI, T. and KANG, E. L. (2010). Fixed rank filtering for spatio-temporal data. *J. Comput. Graph. Statist.* **19** 724–745. With supplementary material available online. [MR2732500](#)
- DAI, F. and XU, Y. (2013). *Approximation Theory and Harmonic Analysis on Spheres and Balls*. Springer, New York, NY.
- DE OLIVEIRA, V., KEDEM, B. and SHORT, D. A. (1997). Bayesian prediction of transformed Gaussian random fields. *J. Amer. Statist. Assoc.* **92** 1422–1433.
- FAN, M. (2015). A note on spherical needlets. Preprint. Available at [arXiv:1508.05406](#).
- FAN, M., PAUL, D., LEE, C. M. T. and MATSUO, T. (2017). Modeling tangential vector fields on a sphere. *J. Amer. Statist. Assoc.* To appear.
- FAN, M., PAUL, D., LEE, C. M. T. and MATSUO, T. (2018). Supplement to “A multi-resolution model for non-Gaussian random fields on a sphere with application to ionospheric electrostatic potentials.” DOI:[10.1214/17-AOAS1104SUPP](#).
- GELMAN, A., ROBERTS, G. O. and GILKS, W. R. (1996). Efficient Metropolis jumping rules. *Bayesian Stat.* **5** 599–608.
- GENTON, M. G. and KLEIBER, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statist. Sci.* **30** 147–163. [MR3353096](#)
- GNEITING, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19** 1327–1349.
- 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.
- GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378.
- GNEITING, T. and RANJAN, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *J. Bus. Econom. Statist.* **29** 411–422.
- GÓRSKI, K. M., HIVON, E., BANDAY, A. J., WANDEL, B. D., HANSEN, F. K., REINECKE, M. and BARTELMANN, M. (2005). HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *Astrophys. J.* **622** 759.
- GUINNESS, J. and FUENTES, M. (2016). Isotropic covariance functions on spheres: Some properties and modeling considerations. *J. Multivariate Anal.* **143** 143–152.
- HEATON, M. J., KLEIBER, W., SAIN, S. R. and WILTBERGER, M. (2015). Emulating and calibrating the multiple-fidelity Lyon–Fedder–Mobarry magnetosphere-ionosphere coupled computer model. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 93–113. [MR3293920](#)
- HUNSUCKER, R. D. and HARGREAVES, J. K. (2007). *The High-Latitude Ionosphere and Its Effects on Radio Propagation*. Cambridge Univ. Press, Cambridge.

- JONES, R. H. (1963). Stochastic processes on a sphere. *Ann. Math. Stat.* **34** 213–218. [MR0170378](#)
- JUN, M. and STEIN, M. L. (2008). Nonstationary covariance models for global data. *Ann. Appl. Stat.* **2** 1271–1289.
- KLEIBER, W., SAIN, S. R., HEATON, M. J., WILTBERGER, M., REESE, C. S., BINGHAM, D. (2013). Parameter tuning for a multi-fidelity dynamical model of the magnetosphere. *Ann. Appl. Stat.* **7** 1286–1310.
- KLEIBER, W., HENDERSHOTT, B., SAIN, S. R. and WILTBERGER, M. (2016). Feature-based validation of the Lyon–Fedder–Mobarry magnetohydrodynamical model. *J. Geophys. Res.* **121** 1192–1200.
- 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.
- LYON, J. G., FEDDER, J. A. and MOBARRY, C. M. (2004). The Lyon–Fedder–Mobarry (LFM) global MHD magnetospheric simulation code. *J. Atmos. Sol.-Terr. Phys.* **66** 1333–1350.
- MARINUCCI, D. and PECCATI, G. (2011). *Random Fields on the Sphere: Representation, Limit Theorems and Cosmological Applications*. Cambridge Univ. Press, Cambridge.
- MATSUO, T. and RICHMOND, A. D. (2008). Effects of high-latitude ionospheric electric field variability on global thermospheric Joule heating and mechanical energy transfer rate. *J. Geophys. Res.* **113** A07309.
- MATSUO, T., RICHMOND, A. D. and HENSEL, K. (2003). High-latitude ionospheric electric field variability and electric potential derived from DE-2 plasma drift measurements: Dependence on IMF and dipole tilt. *J. Geophys. Res.* **108** SIA 1-1–SIA 1-15.
- MATSUO, T., RICHMOND, A. D. and NYCHKA, D. W. (2002). Modes of high-latitude electric field variability derived from DE-2 measurements: Empirical Orthogonal Function (EOF) analysis. *Geophys. Res. Lett.* **29** 11-1–11-4.
- NARCOWICH, F. J., PETRUSHEV, P. and WARD, J. D. (2006). Localized tight frames on spheres. *SIAM J. Math. Anal.* **38** 574–594.
- PALACIOS, M. B. and STEEL, M. F. J. (2006). Non-Gaussian Bayesian geostatistical modeling. *J. Amer. Statist. Assoc.* **101** 604–618. [MR2281244](#)
- PALMROTH, M., JANHUNEN, P., PULKKINEN, T. I., AKSNES, A., LU, G., ØSTGAARD, N., WATTERMANN, J., REEVES, G. D. and GERMANY, G. A. (2005). Assessment of ionospheric Joule heating by GUMICS-4 MHD simulation, AMIE, and satellite-based statistics: Towards a synthesis. *Ann. Geophysicae* **23** 2051–2068.
- PERRON, M. and SURA, P. (2013). Climatology of non-Gaussian atmospheric statistics. *J. Climate* **26** 1063–1083.
- RØISLIEN, J. and OMRE, H. (2006). T-distributed random fields: A parametric model for heavy-tailed well-log data. *Math. Geol.* **38** 821–849.
- RUOHONIEMI, J. M. and BAKER, K. B. (1998). Large-scale imaging of high-latitude convection with Super Dual Auroral Radar Network HF radar observations. *J. Geophys. Res.* **103** 20797–20811.
- SCHOENBERG, I. J. (1942). Positive definite functions on spheres. *Duke Math. J.* **9** 96–108.
- STEIN, M. L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *Ann. Statist.* **16** 55–63.
- STEIN, M. L. (2007). Spatial variation of total column ozone on a global scale. *Ann. Appl. Stat.* **1** 191–210. [MR2393847](#)
- TERDIK, G. (2015). Angular spectra for non-Gaussian isotropic fields. *Braz. J. Probab. Stat.* **29** 833–865.
- WALLIN, J. and BOLIN, D. (2015). Geostatistical modelling using non-Gaussian Matérn fields. *Scand. J. Stat.* **42** 872–890.
- WEIMER, D. R. (1995). Models of high-latitude electric potentials derived with a least error fit of spherical harmonic coefficients. *J. Geophys. Res.* **100** 19595–19607.

- WEST, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74** 646–648. [MR0909372](#)
- WILTBERGER, M., RIGLER, E. J., MERKIN, V. and LYON, J. G. (2017). Structure of high latitude currents in magnetosphere-ionosphere models. *Space Sci. Rev.* **206** 575–598.
- WOLFE, P. J., GODSILL, S. J. and NG, W.-J. (2004). Bayesian variable selection and regularization for time-frequency surface estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 575–589. [MR2088291](#)
- WOMERSLEY, R. S. (2015). Efficient spherical designs with good geometric properties. Available at <http://web.maths.unsw.edu.au/~rsw/Sphere/EffSphDes/>.
- XU, G. and GENTON, M. G. (2017). Tukey g -and- h random fields. *J. Amer. Statist. Assoc.* **112** 1236–1249.
- ZHANG, B., LOTKO, W., WILTBERGER, M. J., BRAMBLES, O. J. and DAMIANO, P. A. (2011). A statistical study of magnetosphere–ionosphere coupling in the Lyon–Fedder–Mobarry global MHD model. *J. Atmos. Sol.-Terr. Phys.* **73** 686–702.

REDUCING STORAGE OF GLOBAL WIND ENSEMBLES WITH STOCHASTIC GENERATORS¹

BY JAEHONG JEONG^{*}, STEFANO CASTRUCCIO[†], PAOLA CRIPPA[†] AND
MARC G. GENTON^{*}

King Abdullah University of Science and Technology^{} and
University of Notre Dame[†]*

Wind has the potential to make a significant contribution to future energy resources. Locating the sources of this renewable energy on a global scale is however extremely challenging, given the difficulty to store very large data sets generated by modern computer models. We propose a statistical model that aims at reproducing the data-generating mechanism of an ensemble of runs via a Stochastic Generator (SG) of global annual wind data. We introduce an evolutionary spectrum approach with spatially varying parameters based on large-scale geographical descriptors such as altitude to better account for different regimes across the Earth's orography. We consider a multi-step conditional likelihood approach to estimate the parameters that explicitly accounts for nonstationary features while also balancing memory storage and distributed computation. We apply the proposed model to more than 18 million points of yearly global wind speed. The proposed SG requires orders of magnitude less storage for generating surrogate ensemble members from wind than does creating additional wind fields from the climate model, even if an effective lossy data compression algorithm is applied to the simulation output.

REFERENCES

- BAKER, A. H., XU, H., DENNIS, J. M., LEVY, M. N., NYCHKA, D., MICKELSON, S. A., EDWARDS, J., VERTENSTEIN, M. and WEGENER, A. (2014). A methodology for evaluating the impact of data compression on climate simulation data. In *Proceedings of the 23rd International Symposium on High-Performance Parallel and Distributed Computing HPDC '14* 203–214. ACM.
- BAKER, A. H., HAMMERLING, D. M., MICKELSON, S. A., XU, H., STOLPE, M. B., NAVEAU, P., SANDERSON, B., EBERT-UPHOFF, I., SAMARASINGHE, S., DE SIMONE, F., CARBONE, F., GENCARELLI, C. N., DENNIS, J. M., KAY, J. E. and LINDSTROM, P. (2016). Evaluating lossy data compression on climate simulation data within a large ensemble. *Geosci. Model Dev.* **9** 4381–4403.
- BANUELOS-RUEDAS, F., ANGELES-CAMACHO, C. and RIOS-MARCUELLO, S. (2011). *Methodologies Used in the Extrapolation of Wind Speed Data at Different Heights and Its Impact in the Wind Energy Resource Assessment in a Region*. INTECH Open Access Publisher.
- BARTHELMIE, R. J. and PRYOR, S. C. (2014). Potential contribution of wind energy to climate change mitigation. *Nature Climate Change* **4** 684–688.

Key words and phrases. Axial symmetry, nonstationarity, spatio-temporal covariance model, sphere, stochastic generator, surface wind speed.

- BETTS, A. K. and RIDGWAY, W. (1988). Coupling of the radiative, convective, and surface fluxes over the equatorial pacific. *J. Atmos. Sci.* **45** 522–536.
- BLYTH, S., GROOMBRIDGE, B., LYSENKO, I., MILES, L. and NEWTON, A. (2002). *Mountain Watch: Environmental Change & Sustainable Developmental in Mountains*. UNEP-WCMC, Cambridge.
- BOLIN, D. and LINDGREN, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Ann. Appl. Stat.* **5** 523–550.
- BRANSTATOR, G. and TENG, H. (2010). Two limits of initial-value decadal predictability in a CGCM. *J. Climate* **23** 6292–6311.
- CASTRUCCIO, S. and GENTON, M. G. (2014). Beyond axial symmetry: An improved class of models for global data. *Stat* **3** 48–55.
- CASTRUCCIO, S. and GENTON, M. G. (2016). Compressing an ensemble with statistical models: An algorithm for global 3D spatio-temporal temperature. *Technometrics* **58** 319–328.
- CASTRUCCIO, S. and GUINNESS, J. (2017). An evolutionary spectrum approach to incorporate large-scale geographical descriptors on global processes. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 329–344. [MR3611690](#)
- CASTRUCCIO, S. and STEIN, M. L. (2013). Global space–time models for climate ensembles. *Ann. Appl. Stat.* **7** 1593–1611. [MR3127960](#)
- CLARKE, J., ALEGRÍA, A. and PORCU, E. (2016). Regularity properties and simulations of Gaussian random fields on the sphere cross time. ArXiv Preprint [arXiv:1611.02851](#).
- COLLINS, M. (2002). Climate predictability on interannual to decadal time scales: The initial value problem. *Clim. Dyn.* **19** 671–692.
- COLLINS, M. and ALLEN, M. R. (2002). Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability. *J. Climate* **15** 3104–3109.
- DAS, B. (2000). Global covariance modeling: A deformation approach to anisotropy. Ph.D. thesis, Univ. Washington.
- DAVIS, P. J. (1979). *Circulant Matrices*. Wiley, New York. [MR0543191](#)
- GASTINEAU, G., LI, L. and LE TREUT, H. (2009). The Hadley and Walker circulation changes in global warming conditions described by idealized atmospheric simulations. *J. Climate* **22** 3993–4013.
- GNEITING, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19** 1327–1349.
- GUINNESS, J. and HAMMERLING, D. (2017). Compression and conditional emulation of climate model output. *J. Amer. Statist. Assoc.* To appear.
- HEATON, M. J., KATZFUSS, M., BERRETT, C. and NYCHKA, D. W. (2014). Constructing valid spatial processes on the sphere using kernel convolutions. *Environmetrics* **25** 2–15.
- HITCZENKO, M. and STEIN, M. L. (2012). Some theory for anisotropic processes on the sphere. *Stat. Methodol.* **9** 211–227.
- JEONG, J., JUN, M. and GENTON, M. G. (2017). Spherical process models for global spatial statistics. *Statist. Sci.* **32** 501–513. [MR3730519](#)
- JEONG, J., CASTRUCCIO, S., CRIPPA, P. and GENTON, M. G. (2018). Supplement to “Reducing storage of global wind ensembles with stochastic generators.” DOI:[10.1214/17-AOAS1105SUPP](#).
- JONES, R. H. (1963). Stochastic processes on a sphere. *Ann. Math. Stat.* **34** 213–218.
- JUN, M. (2011). Non-stationary cross-covariance models for multivariate processes on a globe. *Scand. J. Stat.* **38** 726–747.
- JUN, M. (2014). Matérn-based nonstationary cross-covariance models for global processes. *J. Multivariate Anal.* **128** 134–146.
- JUN, M. and STEIN, M. L. (2007). An approach to producing space–time covariance functions on spheres. *Technometrics* **49** 468–479. [MR2394558](#)
- JUN, M. and STEIN, M. L. (2008). Nonstationary covariance models for global data. *Ann. Appl. Stat.* **2** 1271–1289.

- KAY, J. E., DESER, C., PHILLIPS, A., MAI, A., HANNAY, C., STRAND, G., ARBLASTER, J. M., BATES, S. C., DANABASOGLU, G., EDWARDS, J., HOLLAND, M., KUSHNER, P., LAMARQUE, J.-F., LAWRENCE, D., LINDSAY, K., MIDDLETON, A., MUNOZ, E., NEALE, R., OLESON, K., POLVANI, L. and VERTENSTEIN, M. (2015). The community earth system model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Am. Meteorol. Soc.* **96** 1333–1349.
- LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498.
- LORENZ, E. N. (1963). Deterministic nonperiodic flow. *J. Atmos. Sci.* **20** 130–141.
- MA, C. (2015). Isotropic covariance matrix functions on all spheres. *Math. Geosci.* **47** 699–717.
- MCINNES, K. L., ERWIN, T. A. and BATHOLS, J. M. (2011). Global climate model projected changes in 10 m wind speed and direction due to anthropogenic climate change. *Atmos. Sci. Lett.* **12** 325–333.
- MEARNS, L., HULME, M., CARTER, T., LEEMANS, R., LAL, M. and WHETTON, P. (2001). Climate scenario development. In *Climate Change 2001: The Scientific Basis*. Cambridge Univ. Press, Cambridge.
- MOOMAW, W., YAMBA, F., KAMIMOTO, M., MAURICE, L., NYBOER, J., URAMA, K. and WEIR, T. (2011). Renewable energy and climate change. In *IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation* (O. Edenhofer, R. Pichs-Madruga, Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen and S. Schlömer, eds.) 161–208. Cambridge Univ. Press, Cambridge.
- OBAMA, B. (2017). The irreversible momentum of clean energy. *Science* **355** 126–129.
- POPPICK, A. and STEIN, M. L. (2014). Using covariates to model dependence in nonstationary, high-frequency meteorological processes. *Environmetrics* **25** 293–305.
- PORCU, E., BEVILACQUA, M. and GENTON, M. G. (2016). Spatio-temporal covariance and cross-covariance functions of the great circle distance on a sphere. *J. Amer. Statist. Assoc.* **111** 888–898. [MR3538713](#)
- PRIESTLEY, M. B. (1965). Evolutionary spectra and non-stationary processes. *J. Roy. Statist. Soc. Ser. B* 204–237.
- SOMAN, S. S., ZAREIPOUR, H., MALIK, O. and MANDAL, P. (2010). A review of wind power and wind speed forecasting methods with different time horizons. In *North American Power Symposium (NAPS)*, 2010 1–8. IEEE.
- STEIN, M. L. (2007). Spatial variation of total column ozone on a global scale. *Ann. Appl. Stat.* **1** 191–210. [MR2393847](#)
- TAGLE, F., CASTRUCCIO, S., CRIPPA, P. and GENTON, M. G. (2017). Assessing potential wind energy resources in Saudi Arabia with a skew-t distribution. ArXiv Preprint [arXiv:1703.04312](#).
- 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.
- TUKEY, J. W. (1967). An introduction to the calculations of numerical spectrum analysis. In *Advanced Seminar on Spectral Analysis of Time Series* (B. Harris, ed.) 25–46. Wiley, New York.
- 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.
- VAUGHAN, R. A. and CRACKNELL, A. P. (2013). *Remote Sensing and Global Climate Change* **24**. Springer Science & Business Media.
- WISER, R., YANG, Z., HAND, M., HOHMEYER, O., INFELD, D., JENSEN, P. H., NIKOLAEV, V., O'MALLEY, M., SINDEN, G. and ZERVOS, A. (2011). Wind energy. In *IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation* (O. Edenhofer, R. Pichs-Madruga,

- Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen and S. Schlömer, eds.) 535–608. Cambridge Univ. Press, Cambridge.
- XU, G. and GENTON, M. G. (2017). Tukey g -and- h random fields. *J. Amer. Statist. Assoc.* **112** 1236–1249.
- ZHU, X. and GENTON, M. G. (2012). Short-term wind speed forecasting for power system operations. *Int. Stat. Rev.* **80** 2–23. [MR2990340](#)

MSIQ: JOINT MODELING OF MULTIPLE RNA-SEQ SAMPLES FOR ACCURATE ISOFORM QUANTIFICATION

BY WEI VIVIAN LI^{*,1}, ANQI ZHAO[†], SHIHUA ZHANG^{‡,2} AND
JINGYI JESSICA LI^{*,1,3}

University of California, Los Angeles^{}, Harvard University[†] and
Chinese Academy of Sciences[‡]*

Next-generation RNA sequencing (RNA-seq) technology has been widely used to assess full-length RNA isoform abundance in a high-throughput manner. RNA-seq data offer insight into gene expression levels and transcriptome structures, enabling us to better understand the regulation of gene expression and fundamental biological processes. Accurate isoform quantification from RNA-seq data is challenging due to the information loss in sequencing experiments. A recent accumulation of multiple RNA-seq data sets from the same tissue or cell type provides new opportunities to improve the accuracy of isoform quantification. However, existing statistical or computational methods for multiple RNA-seq samples either pool the samples into one sample or assign equal weights to the samples when estimating isoform abundance. These methods ignore the possible heterogeneity in the quality of different samples and could result in biased and unrobust estimates. In this article, we develop a method, which we call “joint modeling of multiple RNA-seq samples for accurate isoform quantification” (MSIQ), for more robust isoform quantification by integrating multiple RNA-seq samples under a Bayesian framework. Our method aims to (1) identify a consistent group of samples with homogeneous quality and (2) improve isoform quantification accuracy by jointly modeling multiple RNA-seq samples and allowing for higher weights on the consistent group. We show that MSIQ provides a consistent estimator of isoform abundance, and we demonstrate the accuracy and effectiveness of MSIQ compared with alternative methods through simulation studies on *D. melanogaster* genes. We justify MSIQ’s advantages over existing approaches via application studies on real RNA-seq data of human embryonic stem cells, brain tissues, and the HepG2 immortalized cell line. We also perform a comprehensive analysis of how the isoform quantification accuracy would be affected by RNA-seq sample heterogeneity and different experimental protocols.

REFERENCES

- ADAMSKI, M. G., GUMANN, P. and BAIRD, A. E. (2014). A method for quantitative analysis of standard and high-throughput qPCR expression data based on input sample quantity. *PLoS ONE* **9** e103917.

Key words and phrases. Isoform abundance estimation, joint inference from multiple samples, RNA sequencing, Bayesian hierarchical models, Gibbs sampling, data heterogeneity.

- BARRETT, T., WILHITE, S. E., LEDOUX, P., EVANGELISTA, C., KIM, I. F., TOMASHEVSKY, M., MARSHALL, K. A., PHILLIPPY, K. H., SHERMAN, P. M., HOLKO, M. et al. (2013). NCBI GEO: Archive for functional genomics data sets—Update. *Nucleic Acids Res.* **41** D991–D995.
- BEHR, J., KAHLES, A., ZHONG, Y., SREEDHARAN, V. T., DREWE, P. and RÄTSCH, G. (2013). MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples. *Bioinformatics* **29** 2529–2538.
- BERNARD, E., JACOB, L., MAIRAL, J. and VERT, J.-P. (2014). Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. *Bioinformatics* **30** 2447–2455.
- COLLADO-TORRES, L., NELLORE, A., KAMMERS, K., ELLIS, S. E., TAUB, M. A., HANSEN, K. D., JAFFE, A. E., LANGMEAD, B. and LEEK, J. T. (2017). Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35** 319–321.
- CONESA, A., MADRIGAL, P., TARAZONA, S., GOMEZ-CABRERO, D., CERVERA, A., MCPHERSON, A., SZCZEŚNIAK, M. W., GAFFNEY, D. J., ELO, L. L., ZHANG, X. et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17** 13.
- GERMAIN, P.-L., VITRIOLO, A., ADAMO, A., LAISE, P., DAS, V. and TESTA, G. (2016). RNAon-theBENCH: Computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. *Nucleic Acids Res.* **44** 5054–5067.
- GRIEBEL, T., ZACHER, B., RIBECA, P., RAINERI, E., LACROIX, V., GUIGÓ, R. and SAMMETH, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.* **40** 10073–10083.
- HANSEN, K. D., WU, Z., IRIZARRY, R. A. and LEEK, J. T. (2011). Sequencing technology does not eliminate biological variability. *Nat. Biotechnol.* **29** 572–573.
- HARROW, J., FRANKISH, A., GONZALEZ, J. M., TAPANARI, E., DIEKHANS, M., KOKOCINSKI, F., AKEN, B. L., BARRELL, D., ZADISSA, A., SEARLE, S. et al. (2012). GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **22** 1760–1774.
- JIANG, H. and WONG, W. H. (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25** 1026–1032.
- KANITZ, A., GYPAS, F., GRUBER, A. J., GRUBER, A. R., MARTIN, G. and ZAVOLAN, M. (2015). Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.* **16** 1–26.
- KATZ, Y., WANG, E. T., AIROLDI, E. M. and BURGE, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7** 1009–1015.
- KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. and HAUSSLER, D. (2002). The human genome browser at UCSC. *Genome Res.* **12** 996–1006.
- KULKARNI, M. M. (2011). Digital multiplexed gene expression analysis using the NanoString nCounter system. *Curr. Protoc. Mol. Biol.* Unit25B.10.
- LI, B. and DEWEY, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12** 323.
- LI, J. J., JIANG, C.-R., BROWN, J. B., HUANG, H. and BICKEL, P. J. (2011). Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc. Natl. Acad. Sci. USA* **108** 19867–19872.
- LI, W. V., ZHAO, A., ZHANG, S. and LI, J. J. (2018). Supplement to “MSIQ: Joint modeling of multiple RNA-seq samples for accurate isoform quantification.” DOI:[10.1214/17-AOAS1100SUPP](https://doi.org/10.1214/17-AOAS1100SUPP).
- LIN, Y.-Y., DAO, P., HACH, F., BAKHSHI, M., MO, F., LAPUK, A., COLLINS, C. and SAHINNALP, S. C. (2012). CLIQ: Accurate comparative detection and quantification of expressed isoforms in a population. In *Algorithms in Bioinformatics* 178–189. Springer, Berlin.
- MACAULAY, I. C. and VOET, T. (2014). Single cell genomics: Advances and future perspectives. *PLoS Genet.* **10** e1004126.
- MEZLINI, A. M., SMITH, E. J., FIUME, M., BUSKE, O., SAVICH, G. L., SHAH, S., APARICIO, S., CHIANG, D. Y., GOLDENBERG, A. and BRUDNO, M. (2013). iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res.* **23** 519–529.

- MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L. and WOLD, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5** 621–628.
- PACHTER, L. (2011). Models for transcript quantification from RNA-Seq. Preprint. Available at [arXiv:1104.3889](https://arxiv.org/abs/1104.3889).
- PATRO, R., MOUNT, S. M. and KINGSFORD, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32** 462–464.
- PRUITT, K. D., BROWN, G. R., HIATT, S. M., THIBAUD-NISSEN, F., ASTASHYN, A., ERMO-LAEVA, O., FARRELL, C. M., HART, J., LANDRUM, M. J., MCGARVEY, K. M. et al. (2014). RefSeq: An update on mammalian reference sequences. *Nucleic Acids Res.* **42** D756–D763.
- QUAIL, M. A., SMITH, M., COUPLAND, P., OTTO, T. D., HARRIS, S. R., CONNOR, T. R., BERTONI, A., SWERDLOW, H. P. and GU, Y. (2012). A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13** 341.
- ROBERTS, A. and PACHTER, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* **10** 71–73.
- ROSENBLOOM, K. R., ARMSTRONG, J., BARBER, G. P., CASPER, J., CLAWSON, H., DIEKHANS, M., DRESZER, T. R., FUJITA, P. A., GURUVADOO, L., HAEUSSLER, M. et al. (2015). The UCSC genome browser database: 2015 update. *Nucleic Acids Res.* **43** D670–D681.
- ROSSELL, D., ATTOLINI, C. S.-O., KROISS, M. and STÖCKER, A. (2014). Quantifying alternative splicing from paired-end RNA-sequencing data. *Ann. Appl. Stat.* **8** 309.
- SAKHARKAR, M. K., CHOW, V. T. and KANGUEANE, P. (2004). Distributions of exons and introns in the human genome. *In Silico Biol.* **4** 387–393.
- STEIJGER, T., ABRIL, J. F., ENGSTRÖM, P. G., KOKOCINSKI, F., HUBBARD, T. J., GUIGÓ, R., HARROW, J., BERTONE, P., CONSORTIUM, R. et al. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10** 1177–1184.
- TRAPNELL, C., PACHTER, L. and SALZBERG, S. L. (2009). Tophat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25** 1105–1111.
- TRAPNELL, C., WILLIAMS, B. A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M. J., SALZBERG, S. L., WOLD, B. J. and PACHTER, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28** 511–515.
- WANG, Z., GERSTEIN, M. and SNYDER, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10** 57–63.
- WU, A. R., NEFF, N. F., KALISKY, T., DALERBA, P., TREUTLEIN, B., ROTHENBERG, M. E., MBURU, F. M., MANTALAS, G. L., SIM, S., CLARKE, M. F. et al. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11** 41–46.
- YE, Y. and LI, J. J. (2016). NMFP: A non-negative matrix factorization based preselection method to increase accuracy of identifying mRNA isoforms from RNA-seq data. *BMC Genomics* **17** 127.
- ZHANG, J., KUO, C.-C. J. and CHEN, L. (2014). WemIQ: An accurate and robust isoform quantification method for RNA-seq data. *Bioinformatics* **31** 878–885.

KERNEL-PENALIZED REGRESSION FOR ANALYSIS OF MICROBIOME DATA

BY TIMOTHY W. RANDOLPH^{*,1,2}, SEN ZHAO^{†,3}, WADE COPELAND^{*},
MEREDITH HULLAR^{*,1} AND ALI SHOJAIE^{†,2,3}

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

The analysis of human microbiome data is often based on dimension-reduced graphical displays and clusterings derived from vectors of microbial abundances in each sample. Common to these ordination methods is the use of biologically motivated definitions of similarity. Principal coordinate analysis, in particular, is often performed using ecologically defined distances, allowing analyses to incorporate context-dependent, non-Euclidean structure. In this paper, we go beyond dimension-reduced ordination methods and describe a framework of high-dimensional regression models that extends these distance-based methods. In particular, we use kernel-based methods to show how to incorporate a variety of extrinsic information, such as phylogeny, into penalized regression models that estimate taxon-specific associations with a phenotype or clinical outcome. Further, we show how this regression framework can be used to address the compositional nature of multivariate predictors comprised of relative abundances; that is, vectors whose entries sum to a constant. We illustrate this approach with several simulations using data from two recent studies on gut and vaginal microbiomes. We conclude with an application to our own data, where we also incorporate a significance test for the estimated coefficients that represent associations between microbial abundance and a percent fat.

REFERENCES

- AITCHISON, J. (1982). The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B* **44** 139–177. With discussion. [MR0676206](#)
- AITCHISON, J. (2003a). A concise guide to compositional data analysis. In *2nd Compositional Data Analysis Workshop*.
- AITCHISON, J. (2003b). *The Statistical Analysis of Compositional Data*. The Blackburn Press, Caldwell, NJ.
- ANDERSON, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62** 245–253, 320. [MR2226579](#)
- BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. [MR1869245](#)
- BÜHLMANN, P., KALISCH, M. and MEIER, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annu. Rev. Statist. Appl.* **1** 255–278.
- CAI, T. T. and HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179.

Key words and phrases. Compositional data, distance-based analysis, kernel methods, microbial community data, penalized regression.

- CHEN, J., BITTINGER, K., CHARLSON, E. S., HOFFMANN, C., LEWIS, J., WU, G. D., COLL-MAN, R. G., BUSHMAN, F. D. and LI, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* **28** 2106–2113.
- CLAESSON, M. J., JEFFERY, I. B., CONDE, S., POWER, S. E., O’CONNOR, E. M., CUSACK, S., HARRIS, H. M., COAKLEY, M., LAKSHMINARAYANAN, B., O’SULLIVAN, O., FITZGERALD, G. F., DEANE, J., O’CONNOR, M., HARNEDY, N., O’CONNOR, K., O’MAHONY, D., VAN SINDEREN, D., WALLACE, M., BRENNAN, L., STANTON, C., MARCHESI, J. R., FITZGERALD, A. P., SHANAHAN, F., HILL, C., ROSS, R. P. and O’TOOLE, P. W. (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488** 178–184.
- EGOZCUE, J. J. and PAWLOWSKY-GLAHN, V. (2011). Basic concepts and procedures. In *Compositional Data Analysis: Theory and Applications* (V. Pawlowsky-Glahn and A. Buccianti, eds.) 12–28. Wiley, Chichester.
- EVANS, S. N. and MATSEN, F. A. (2012). The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 569–592.
- FRANKLIN, J. N. (1978). Minimum principles for ill-posed problems. *SIAM J. Math. Anal.* **9** 638–650.
- FREYTAG, S., MANITZ, J., SCHLATHER, M., KNEIB, T., AMOS, C. I., RISCH, A., CHANG-CLAUDE, J., HEINRICH, J. and BICKEBÖLLER, H. (2013). A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. *Hum. Hered.* **76** 64–75.
- FRIEDMAN, J. and ALM, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8** e1002687.
- FUKUYAMA, J., MCMURDIE, P. J., DETHLEFSEN, L., RELMAN, D. A. and HOLMES, S. (2012). Comparisons of distance methods for combining covariates and abundances in microbiome studies. In *Pacific Symposium on Biocomputing 2012* 213–224.
- GOLUB, G. H. and VAN LOAN, C. F. (2012). *Matrix Computations*. Johns Hopkins Univ. Press, Baltimore, MD.
- GOODRICH, J. K., RIENZI, S. C. D., POOLE, A. C., KOREN, O., WALTERS, W. A., CAPO-RASO, J. G., KNIGHT, R. and LEY, R. E. (2014). Conducting a microbiome study. *Cell* **158** 250–262.
- GOWER, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53** 325–338.
- GREENACRE, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, Cambridge, MA.
- GRETTON, A., HERBRICH, R., SMOLA, A., BOUSQUET, O. and SCHÖLKOPF, B. (2005). Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6** 2075–2129.
- HAMADY, M. and KNIGHT, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res.* **19** 1141–1152.
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.
- HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.
- HULLAR, M. A., LANCASTER, S. M., LI, F., TSENG, E., BEER, K., ATKINSON, C., WÄHÄLÄ, K., COPELAND, W. K., RANDOLPH, T. W., NEWTON, K. M. and LAMPE, J. W. (2015). Enterolignan-producing phenotypes are associated with increased gut microbial diversity and altered composition in premenopausal women in the United States. *Cancer Epidemiol. Biomark. Prev.* **24** 546–554.
- JOSSE, J. and HOLMES, S. (2016). Measuring multivariate association and beyond. *Stat. Surv.* **10** 132–167.

- KIM, S. and XING, E. P. (2010). Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)* (J. Fürnkranz and T. Joachims, eds.) 543–550.
- KOREN, O., KNIGHTS, D., GONZALEZ, A., WALDRON, L., SEGATA, N., KNIGHT, R., HUTTENHOWER, C. and LEY, R. E. (2013). A guide to enterotypes across the human body: Meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* **9** e1002863.
- KUCZYNSKI, J., LIU, Z., LOZUPONE, C., McDONALD, D., FIERER, N. and KNIGHT, R. (2010). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods* **7** 813–819.
- KURTZ, Z. D., MUELLER, C. L., MIRALDI, E. R., LITTMAN, D. R., BLASER, M. J. and BONNEAU, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11** e1004226.
- LI, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Statist. Appl.* **2** 73–94.
- LI, C. and LI, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24** 1175–1182.
- LOVELL, D., PAWLOWSKY-GLAHN, V., EGOZCUE, J. J., MARGUERAT, S. and BÄHLER, J. (2015). Proportionality: A valid alternative to correlation for relative data. *PLoS Comput. Biol.* **11** e1004075.
- LOZUPONE, C. and KNIGHT, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **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. *Appl. Environ. Microbiol.* **73** 1576–1585.
- MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1980). *Multivariate Analysis*. Academic Press, San Diego, CA.
- 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* **8** e56859.
- PAN, W. (2011). Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet. Epidemiol.* **35** 211–216.
- 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.
- PEARSON, K. (1896). Mathematical contributions to the theory of evolution—On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. R. Soc. Lond.* **60** 489–498.
- PEKALSKA, E., PACLIK, P. and DUIN, R. P. (2002). A generalized kernel approach to dissimilarity-based classification. *J. Mach. Learn. Res.* **2** 175–211.
- PURDOM, E. (2011). Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree. *Ann. Appl. Stat.* **5** 2326–2358.
- RANDOLPH, T. W., HAREZLAK, J. and FENG, Z. (2012). Structured penalties for functional linear models—Partially empirical eigenvectors for regression. *Electron. J. Stat.* **6** 323–353.
- RANDOLPH, W. T., ZHAO, S., COPELAND, W., HULLAR, M. and SHOJAIE, A. (2018). Supplement to “Kernel-penalized regression for analysis of microbiome data.” DOI:[10.1214/17-AOAS1102SUPP](https://doi.org/10.1214/17-AOAS1102SUPP).
- ROBERT, P. and ESCOUFIER, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV -coefficient. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **25** 257–265. [MR0440801](https://doi.org/10.1093/bjstat/25.3.257)
- RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge Univ. Press, New York.

- SCHAID, D. J. (2010). Genomic similarity and kernel methods I: Advancements by building on mathematical and statistical foundations. *Hum. Hered.* **70** 109–131.
- SCHIFANO, E. D., EPSTEIN, M. P., BIELAK, L. F., JHUN, M. A., KARDIA, S. L. R., PEYSER, P. A. and LIN, X. (2012). SNP set association analysis for familial data. *Genet. Epidemiol.* **36** 797–810.
- SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.
- SRINIVASAN, S., HOFFMAN, N. G., MORGAN, M. T., MATSEN, F. A., FIEDLER, T. L., HALL, R. W., ROSS, F. J., MCCOY, C. O., BUMGARNER, R., MARRAZZO, J. M. et al. (2012). Bacterial communities in women with bacterial vaginosis: High resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLoS ONE* **7** e37818.
- SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898.
- SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265.
- TANASEICHUK, O., BORNEMAN, J. and JIANG, T. (2014). Phylogeny-based classification of microbial communities. *Bioinformatics* **30** 449–456.
- THE HUMAN MICROBIOME PROJECT CONSORTIUM (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486** 207–214.
- TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. [MR2850205](#)
- TOLOSANA-DELGADO, V. and VAN DEN BOOGART, K. G. (2011). Linear models with compositions in R. In *Compositional Data Analysis: Theory and Applications* (V. Pawlowsky-Glahn and A. Buccianti, eds.) 12–28. Wiley, Chichester.
- VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202.
- VAN LOAN, C. F. (1976). Generalizing the singular value decomposition. *SIAM J. Numer. Anal.* **13** 76–83.
- YATSUNENKO, T., REY, F. E., MANARY, M. J., TREHAN, I., DOMINGUEZ-BELLO, M. G., CONTRERAS, M., MAGRIS, M., HIDALGO, G., BALDASSANO, R. N., ANOKHIN, A. P., HEATH, A. C., WARNER, B., REEDER, J., KUCZYNSKI, J., CAPORASO, J. G., LOZUPONE, C. A., LAUBER, C., CLEMENTE, J. C., KNIGHTS, D., KNIGHT, R. and GORDON, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature* **486** 222–227.
- ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. [MR3153940](#)
- ZHAO, S. and SHOJAIE, A. (2016). A significance test for graph-constrained estimation. *Biometrics* **72** 484–493. [MR3515775](#)
- ZHAO, N., CHEN, J., CARROLL, I. M., RINGEL-KULKA, T., EPSTEIN, M. P., ZHOU, H., ZHOU, J. J., RINGEL, Y., LI, H. and WU, M. C. (2015). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.* **96** 797–807.

POWERFUL TEST BASED ON CONDITIONAL EFFECTS FOR GENOME-WIDE SCREENING¹

BY YAOWU LIU AND JUN XIE

Harvard University and Purdue University

This paper considers testing procedures for screening large genome-wide data, where we examine hundreds of thousands of genetic variants, for example, single nucleotide polymorphisms (SNP), on a quantitative phenotype. We screen the whole genome by SNP sets and propose a new test that is based on conditional effects from multiple SNPs. The test statistic is developed for weak genetic effects and incorporates correlations among genetic variables, which may be very high due to linkage disequilibrium. The limiting null distribution of the test statistic and the power of the test are derived. Under appropriate conditions, the test is shown to be more powerful than the minimum p -value method, which is based on marginal SNP effects and is the most commonly used method in genome-wide screening. The proposed test is also compared with other existing methods, including the Higher Criticism (HC) test and the sequence kernel association test (SKAT), through simulations and analysis of a real genome data set. For typical genome-wide data, where effects of individual SNPs are weak and correlations among SNPs are high, the proposed test is more advantageous and clearly outperforms the other methods in the literature.

REFERENCES

- ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. [MR2906877](#)
- BALLARD, D. H., CHO, J. and ZHAO, H. (2010). Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genet. Epidemiol.* **34** 201–212.
- 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](#)
- CAI, T., LIU, W. and XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 349–372.
- CHEN, B. E., SAKODA, L. C., HSING, A. W. and ROSENBERG, P. S. (2006). Resampling-based multiple hypothesis testing procedures for genetic case-control association studies. *Genet. Epidemiol.* **30** 495–507.
- CUI, J., STAHL, E. A., SAEVARSDOTTIR, S., MICELI, C., DIOGO, D., TRYNKA, G., RAJ, T., MIRKOV, M. U., CANHAO, H., IKARI, K. et al. (2013). Genome-wide association study and gene expression analysis identifies CD84 as a predictor of response to etanercept therapy in rheumatoid arthritis. *PLoS Genet.* **9** e1003394.
- DOLCINO, M., OTTRIA, A., BARBIERI, A., PATUZZO, G., TINAZZI, E., ARGENTINO, G., BERI, R., LUNARDI, C. and PUCCETTI, A. (2015). Gene expression profiling in peripheral blood cells and synovial membranes of patients with psoriatic arthritis. *PLoS ONE* **10** e0128262.

Key words and phrases. Asymptotically powerful, high dimensional test, limiting null distribution.

- DONOHO, D. and JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.* **30** 1–25. [MR3317751](#)
- FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911.
- GOEMAN, J. J., VAN DE GEER, S. A. and VAN HOUWELINGEN, H. C. (2006). Testing against a high dimensional alternative. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 477–493.
- HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. [MR2662357](#)
- INGSTER, Y. I., TSYBAKOV, A. B., VERZELEN, N. et al. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526.
- LEE, S., ABECASIS, G. R., BOEHNKE, M. and LIN, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *Am. J. Hum. Genet.* **95** 5–23.
- LEISCH, F., WEINGESSEL, A. and HORNIK, K. (1998). On the generation of correlated artificial binary data.
- LI, J. and ZHONG, P. (2017). A rate optimal procedure for recovering sparse differences between high-dimensional means under dependence. *Ann. Statist.* **45** 557–590. [MR3650393](#)
- LIU, Y. and XIE, J. (2018). Supplement to “Powerful test based on conditional effects for genome-wide screening.” DOI:[10.1214/17-AOAS1103SUPP](#).
- LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. [MR3210970](#)
- PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.
- SHAM, P. C. and PURCELL, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15** 335–346.
- TAYLOR, J., LOFTUS, J. and TIBSHIRANI, R. (2013). Tests in adaptive regression via the Kac-Rice formula. Preprint. Available at [arXiv:1308.3020](#).
- WU, M. C., KRAFT, P., EPSTEIN, M. P., TAYLOR, D. M., CHANOCK, S. J., HUNTER, D. J. and LIN, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.* **86** 929–942.
- WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89** 82–93.
- WU, Z., SUN, Y., HE, S., CHO, J., ZHAO, H. and JIN, J. (2014). Detection boundary and higher criticism approach for rare and weak genetic effects. *Ann. Appl. Stat.* **8** 824–851. [MR3262536](#)

FAST INFERENCE OF INDIVIDUAL ADMIXTURE COEFFICIENTS USING GEOGRAPHIC DATA¹

BY KEVIN CAYE^{*}, FLORA JAY[†], OLIVIER MICHEL^{*} AND
OLIVIER FRANÇOIS^{*,2}

Université Grenoble-Alpes^{} and Université Paris-Sud[†]*

Accurately evaluating the distribution of genetic ancestry across geographic space is one of the main questions addressed by evolutionary biologists. This question has been commonly addressed through the application of Bayesian estimation programs allowing their users to estimate individual admixture proportions and allele frequencies among putative ancestral populations. Following the explosion of high-throughput sequencing technologies, several algorithms have been proposed to cope with computational burden generated by the massive data in those studies. In this context, incorporating geographic proximity in ancestry estimation algorithms is an open statistical and computational challenge. In this study, we introduce new algorithms that use geographic information to estimate ancestry proportions and ancestral genotype frequencies from population genetic data. Our algorithms combine matrix factorization methods and spatial statistics to provide estimates of ancestry matrices based on least-squares approximation. We demonstrate the benefit of using spatial algorithms through extensive computer simulations, and we provide an example of application of our new algorithms to a set of spatially referenced samples for the plant species *Arabidopsis thaliana*. Without loss of statistical accuracy, the new algorithms exhibit runtimes that are much shorter than those observed for previously developed spatial methods. Our algorithms are implemented in the R package, *tess3r*.

REFERENCES

- 1000 GENOMES PROJECT CONSORTIUM, AUTON, A., BROOKS, L. D., DURBIN, R. M., GARRISON, E. P., KANG, H. M., KORBEL, J. O., MARCHINI, J. L., MCCARTHY, S., MCVEAN, G. A. and ABECASIS, G. R. (2015). A global reference for human genetic variation. *Nature* **526** 68–74.
- ALEXANDER, D. H. and LANGE, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinform.* **12** 246.
- BARAN, Y., QUINTELA, I., CARRACEDO, Á., PASANIUC, B. and HALPERIN, E. (2013). Enhanced localization of genetic samples through linkage-disequilibrium correction. *Am. J. Hum. Genet.* **92** 882–894.
- BELKIN, M. and NIYOGI, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **6** 1373–1396.
- 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](#)
- BERTSEKAS, D. P. (1995). *Nonlinear Programming*. Athena Scientific, Nashua, USA.

Key words and phrases. Ancestry estimation algorithms, genotypic data, geographic data, fast algorithms.

- BRADBURY, G. S., RALPH, P. L. and COOP, G. M. (2016). A spatial framework for understanding population structure and admixture. *PLoS Genet.* **12** e1005703.
- CAI, D., HE, X., HAN, J. and HUANG, T. S. (2011). Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 1548–1560.
- CARBON, S., IRELAND, A., MUNGALL, C. J., SHU, S., MARSHALL, B., LEWIS, S., AMIGO HUB and WEB PRESENCE WORKING GROUP (2009). AmiGO: Online access to ontology and annotation data. *Bioinformatics* **25** 288–289.
- CAVALLI, L. L., MENOZZI, P. and PIAZZA, A. (1994). *The History and Geography of Human Genes*. Princeton Univ. Press, Princeton, NJ.
- CAYE, K., DEIST, T. M., MARTINS, H., MICHEL, O. and FRANÇOIS, O. (2016). TESS3: Fast inference of spatial population structure and genome scans for selection. *Mol. Ecol. Resour.* **16** 540–548.
- CHEN, C., DURAND, E., FORBES, F. and FRANÇOIS, O. (2007). Bayesian clustering algorithms ascertaining spatial population structure: A new computer program and a comparison study. *Mol. Ecol. Notes* **7** 747–756.
- CICHOCKI, A., ZDUNEK, R., PHAN, A. H. and AMARI, S. I. (2009). *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. Wiley, Chichester.
- CORANDER, J., SIRÉN, J. and ARJAS, E. (2008). Bayesian spatial modeling of genetic population structure. *Comput. Statist.* **23** 111–129. [MR2434757](#)
- CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.
- DEVLIN, B. and ROEDER, K. (1999). Genomic control for association studies. *Biometrics* **55** 997–1004.
- DURAND, E., JAY, F., GAGGIOTTI, O. E. and FRANÇOIS, O. (2009). Spatial inference of admixture proportions and secondary contact zones. *Mol. Biol. Evol.* **26** 1963–1973.
- EASTMENT, H. and KRZANOWSKI, W. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics* **24** 73–77.
- EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. [MR2054289](#)
- ENGELHARDT, B. E. and STEPHENS, M. (2010). Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. *PLoS Genet.* **6** e1001117.
- EPPELSON, B. K. (2003). *Geographical Genetics*. Princeton Univ. Press, Princeton, NJ.
- EPPELSON, B. K. and LI, T. (1996). Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. *Proc. Natl. Acad. Sci. USA* **93** 10528–10532.
- FOURNIER-LEVEL, A., KORTE, A., COOPER, M. D., NORDBORG, M., SCHMITT, J. and WILCZEK, A. M. (2011). A map of local adaptation in *Arabidopsis thaliana*. *Science* **334** 86–89.
- FRANÇOIS, O. and DURAND, E. (2010). Spatially explicit Bayesian clustering models in population genetics. *Mol. Ecol. Resour.* **10** 773–784.
- FRANÇOIS, O. and WAITS, L. P. (2016). Clustering and assignment methods in landscape genetics. 114–128. Wiley, Chichester.
- FRANÇOIS, O., MARTINS, H., CAYE, K. and SCHOVILLE, S. D. (2016). Controlling false discoveries in genome scans for selection. *Mol. Ecol.* **25** 454–469.
- FRICHOT, E. and FRANÇOIS, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods Ecol. Evol.* **6** 925–929.
- FRICHOT, E., MATHIEU, F., TROUILLON, T., BOUCHARD, G. and FRANÇOIS, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196** 973–983.
- GRIPPO, L. and SCIANDRONE, M. (2000). On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Oper. Res. Lett.* **26** 127–136.

- HANCOCK, A. M., BRACHI, B., FAURE, N., HORTON, M. W., JARYMOWYCZ, L. B., SPERONE, F. G., TOOMAJIAN, C., ROUX, F. and BERGELSON, J. (2011). Adaptation to climate across the *Arabidopsis thaliana* genome. *Science* **334** 83–86.
- HARDY, O. J. and VEKEMANS, X. (1999). Isolation by distance in a continuous population: Reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* **83** 145–154.
- HORTON, M. W., HANCOCK, A. M., HUANG, Y. S., TOOMAJIAN, C., ATWELL, S., AUTON, A., MULIYATI, N. W., PLATT, A., SPERONE, F. G., VILHJÁLMSSON, B. J., NORDBORG, M., BOREVITZ, J. O. and BERGELSON, J. (2012). Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.* **44** 212–216.
- HUDSON, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18** 337–338.
- KIM, J. and PARK, H. (2011). Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM J. Sci. Comput.* **33** 3261–3281.
- KORNELIUSSEN, T. S., ALBRECHTSEN, A. and NIELSEN, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinform.* **15** 356.
- LAO, O., LIU, F., WOLLSTEIN, A. and KAYSER, M. (2014). GAGA: A new algorithm for genomic inference of geographic ancestry reveals fine level population substructure in Europeans. *PLoS Comput. Biol.* **10** e1003480.
- LEE, D. D. and SEUNG, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401** 788–791.
- LI, G. and ZHU, H. (2013). Genetic studies: The linear mixed models in genome-wide association studies. *Open Bioinform. J.* **7** 27–33. [MR3172942](#)
- MALÉCOT, G. (1948). *Les Mathématiques de L'Hérédité*. Masson et Cie., Paris. [MR0027490](#)
- MANTEL, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27** 209–220.
- MARTINS, H., CAYE, K., LUU, K., BLUM, M. G. B. and FRANÇOIS, O. (2016). Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *Mol. Ecol.* **25** 5029–5042.
- POPESCU, A. A., HARPER, A. L., TRICK, M., BANCROFT, I. and HUBER, K. T. (2014). A novel and fast approach for population structure inference using kernel-PCA and optimization. *Genetics* **198** 1421–1431.
- PRITCHARD, J. K., STEPHENS, M. and DONNELLY, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155** 945–959.
- RAJ, A., STEPHENS, M. and PRITCHARD, J. K. (2014). FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197** 573–589.
- RAÑOLA, J. M., NOVEMBRE, J. and LANGE, K. (2014). Fast spatial ancestry via flexible allele frequency surfaces. *Bioinformatics* **30** 2915–2922.
- R CORE TEAM (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at <https://www.R-project.org/>.
- SCHRAIBER, J. G. and AKEY, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nat. Rev. Genet.* **16** 727–740.
- SEGELBACHER, G., CUSHMAN, S. A., EPPERSON, B. K., FORTIN, M. J., FRANCOIS, O., HARDY, O. J., HOLDEREGGER, R., TABERLET, P., WAITS, L. P. and MANEL, S. (2010). Applications of landscape genetics in conservation biology: Concepts and challenges. *Conserv. Genet.* **11** 375–385.
- TANG, H., PENG, J., WANG, P. and RISCH, N. J. (2005). Estimation of individual admixture: Analytical and study design considerations. *Genet. Epidemiol.* **28** 289–301.
- WANG, J. (2017). The computer program structure for assigning individuals to populations: Easy to use but easier to misuse. *Mol. Ecol. Resour.* **17** 981–990.

- WEIR, B. S. (1996). *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*, Vol. 2. Sinauer Associates, Sunderland, MA.
- WOLD, S. (1978). Cross-validated estimation of the number of components in factor and principal components models. *Technometrics* **20** 397–405.
- WOLLSTEIN, A. and LAO, O. (2015). Detecting individual ancestry in the human genome. *Invest. Genet.* **6** 7.
- WRIGHT, S. (1943). Isolation by distance. *Genetics* **28** 114–138.
- YANG, W.-Y., PLATT, A., CHIANG, C. W.-K., ESKIN, E., NOVEMBRE, J. and PASANIUC, B. (2014). Spatial localization of recent ancestors for admixed individuals. *Genes Genomes Genet.* **4** 2505–2518.

A UNIFIED STATISTICAL FRAMEWORK FOR SINGLE CELL AND BULK RNA SEQUENCING DATA

BY LINGXUE ZHU*, JING LEI*,¹, BERNIE DEVLIN^{†,2} AND
KATHRYN ROEDER*,^{2,3}

Carnegie Mellon University and University of Pittsburgh[†]*

Recent advances in technology have enabled the measurement of RNA levels for individual cells. Compared to traditional tissue-level bulk RNA-seq data, single cell sequencing yields valuable insights about gene expression profiles for different cell types, which is potentially critical for understanding many complex human diseases. However, developing quantitative tools for such data remains challenging because of high levels of technical noise, especially the “dropout” events. A “dropout” happens when the RNA for a gene fails to be amplified prior to sequencing, producing a “false” zero in the observed data. In this paper, we propose a Unified RNA-Sequencing Model (URSM) for both single cell and bulk RNA-seq data, formulated as a hierarchical model. URSM borrows the strength from both data sources and carefully models the dropouts in single cell data, leading to a more accurate estimation of cell type specific gene expression profile. In addition, URSM naturally provides inference on the dropout entries in single cell data that need to be imputed for downstream analyses, as well as the mixing proportions of different cell types in bulk samples. We adopt an empirical Bayes’ approach, where parameters are estimated using the EM algorithm and approximate inference is obtained by Gibbs sampling. Simulation results illustrate that URSM outperforms existing approaches both in correcting for dropouts in single cell data, as well as in deconvolving bulk samples. We also demonstrate an application to gene expression data on fetal brains, where our model successfully imputes the dropout genes and reveals cell type specific expression patterns.

REFERENCES

- ABBAS, A. R., WOLSLEGEL, K., SESHASAYEE, D., MODRUSAN, Z. and CLARK, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE* **4** e6098.
- BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. [MR3671776](#)
- BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.
- BRENNECKE, P., ANDERS, S., KIM, J. K., KOŁODZIEJCZYK, A. A., ZHANG, X., PROSERPIO, V., BAYING, B., BENES, V., TEICHMANN, S. A., MARIONI, J. C. et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10** 1093–1095.

Key words and phrases. Single cell RNA sequencing, hierarchical model, empirical Bayes, Gibbs sampling, EM algorithm.

- CAMP, J. G., BADSHA, F., FLORIO, M., KANTON, S., GERBER, T., WILSCH-BRÄUNINGER, M., LEWITUS, E., SYKES, A., HEVERS, W., LANCASTER, M. et al. (2015). Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc. Natl. Acad. Sci. USA* **112** 15672–15677.
- CASELLA, G. and GEORGE, E. I. (1992). Explaining the Gibbs sampler. *Amer. Statist.* **46** 167–174. [MR1183069](#)
- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. With discussion. [MR0501537](#)
- DONOHO, D. and STODDEN, V. (2003). When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*.
- DUPUY, C. and BACH, F. (2016). Online but accurate inference for latent variable models with local Gibbs sampling. *J. Mach. Learn. Res.* **1**.
- FAN, H. C., FU, G. K. and FODOR, S. P. A. (2015). Combinatorial labeling of single cells for gene expression cytometry. *Science* **347** 1258367.
- FINAK, G., MCDAVID, A., YAJIMA, M., DENG, J., GERSUK, V., SHALEK, A. K., SLICHTER, C. K., MILLER, H. W., MCEL RATH, M. J., PRLIC, M., LINSLEY, P. S. and GOT TARD, R. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16** 278.
- FROMER, M., ROUSSOS, P., SIEBERTS, S. K., JOHNSON, J. S., KAVANAGH, D. H., PERU MAL, T. M., RUDERFER, D. M., OH, E. C., TOPOL, A. ET AL. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19** 1442–1453.
- GAUJOUX, R. and SEOIGHE, C. (2012). Semi-supervised nonnegative matrix factorization for gene expression deconvolution: A case study. *Infect. Genet. Evol.* **12** 913–921.
- GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. [MR1141740](#)
- GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.
- GRIFFITHS, T. L. and STEYVERS, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **101** 5228–5235.
- GRÜN, D., LYUBIMOVA, A., KESTER, L., WIEBRANDS, K., BASAK, O., SASAKI, N., CLEVERS, H. and VAN OUDENAARDEN, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525** 251–255.
- GTEX CONSORTIUM (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45** 580–585.
- HAQUE, A., ENGEL, J., TEICHMANN, S. A. and LÖNNBERG, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Gen. Med.* **9** 75.
- HUANG, W.-C., FERRIS, E., CHENG, T., HÖRNDLI, C. S., GLEASON, K., TAMMINGA, C., WAGNER, J. D., BOUCHER, K. M., CHRISTIAN, J. L. and GREGG, C. (2017a). Diverse non-genetic, allele-specific expression effects shape genetic architecture at the cellular level in the mammalian brain. *Neuron* **93** 1094–1109.e7.
- HUANG, M., WANG, J., TORRE, E., DUECK, H., SHAFFER, S., BONASIO, R., MURRAY, J., RAJ, A., LI, M. and ZHANG, N. R. (2017b). Gene expression recovery for single cell RNA sequencing. *BioRxiv*. DOI:10.1101/138677.
- JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.
- KANG, H. J., KAWASAWA, Y. I., CHENG, F., ZHU, Y., XU, X., LI, M., SOUSA, A. M., PLETIKOS, M., MEYER, K. A., SEDMAK, G. et al. (2011). Spatio-temporal transcriptome of the human brain. *Nature* **478** 483–489.
- KHARCHENKO, P. V., SILBERSTEIN, L. and SCADDEN, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11** 740–742.

- KOLODZIEJCZYK, A. A., KIM, J. K., SVENSSON, V., MARIONI, J. C. and TEICHMANN, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58** 610–620.
- LEE, D. D. and SEUNG, H. S. (2000). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems* **13** 556–562. MIT Press, Cambridge, MA.
- 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** 59.
- NEWMAN, A. M., LIU, C. L., GREEN, M. R., GENTLES, A. J., FENG, W., XU, Y., HOANG, C. D., DIEHN, M. and ALIZADEH, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12** 453–457.
- PADOVAN-MERHAR, O. and RAJ, A. (2013). Using variability in gene expression as a tool for studying gene regulation. *Wiley Interdiscip. Rev., Syst. Biol. Med.* **5** 751–759.
- PIERSON, E. and YAU, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16** 241.
- 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](#)
- PRABHAKARAN, S., AZIZI, E. and PE’ER, D. (2016). Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *Proceedings of the 33rd International Conference on Machine Learning* 1070–1079.
- REPSILBER, D., KERN, S., TELAAR, A., WALZL, G., BLACK, G. F., SELBIG, J., PARIDA, S. K., KAUFMANN, S. H. E. and JACOBSEN, M. (2010). Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach. *BMC Bioinform.* **11** 27.
- SATIJA, R., FARRELL, J. A., GENNERT, D., SCHIER, A. F. and REGEV, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33** 495–502.
- SHEN-ORR, S. S., TIBSHIRANI, R., KHATRI, P., BODIAN, D. L., STAEDTLER, F., PERRY, N. M., HASTIE, T., SARWAL, M. M., DAVIS, M. M. and BUTTE, A. J. (2010). Cell type-specific gene expression differences in complex tissues. *Nat. Methods* **7** 287–289.
- SUNKIN, S. M., NG, L., LAU, C., DOLBEARE, T., GILBERT, T. L., THOMPSON, C. L., HAWRYLYCZ, M. and DANG, C. (2013). Allen brain atlas: An integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.* **41** D996–D1008.
- TRAPNELL, C., CACCHIARELLI, D., GRIMSBY, J., POKHAREL, P., LI, S., MORSE, M., LENNON, N. J., LIVAK, K. J., MIKKELSEN, T. S. and RINN, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32** 381–386.
- VALLEJOS, C. A., MARIONI, J. C. and RICHARDSON, S. (2015). BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput. Biol.* **11** e1004333.
- VALLEJOS, C. A., RICHARDSON, S. and MARIONI, J. C. (2016). Beyond comparisons of means: Understanding changes in gene expression at the single-cell level. *Genome Biol.* **17** 1.
- VALLEJOS, C. A., RISSO, D., SCIALDONE, A., DUDOIT, S. and MARIONI, J. C. (2017). Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nat. Methods* **14** 565–571.
- VU, T. N., WILLS, Q. F., KALARI, K. R., NIU, N., WANG, L., RANTALAINEN, M. and PAWITAN, Y. (2016). Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32** 2128–35.
- WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* **1** 1–305.
- ZHONG, Y., WAN, Y.-W., PANG, K., CHOW, L. M. and LIU, Z. (2013). Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinform.* **14** 1.
- ZHU, L., LEI, J., DEVLIN, B. and ROEDER, K. (2018). Supplement to “A unified statistical framework for single cell and bulk RNA sequencing data.” DOI:10.1214/17-AOAS1110SUPP.

MODELING AND ESTIMATION FOR SELF-EXCITING SPATIO-TEMPORAL MODELS OF TERRORIST ACTIVITY

BY NICHOLAS J. CLARK¹ AND PHILIP M. DIXON

Iowa State University

Spatio-temporal hierarchical modeling is an extremely attractive way to model the spread of crime or terrorism data over a given region, especially when the observations are counts and must be modeled discretely. The spatio-temporal diffusion is placed, as a matter of convenience, in the process model allowing for straightforward estimation of the diffusion parameters through Bayesian techniques. However, this method of modeling does not allow for the existence of self-excitation, or a temporal data model dependency, that has been shown to exist in criminal and terrorism data. In this manuscript we will use existing theories on how violence spreads to create models that allow for both spatio-temporal diffusion in the process model as well as temporal diffusion, or self-excitation, in the data model. We will further demonstrate how Laplace approximations similar to their use in Integrated Nested Laplace Approximation can be used to quickly and accurately conduct inference of self-exciting spatio-temporal models allowing practitioners a new way of fitting and comparing multiple process models. We will illustrate this approach by fitting a self-exciting spatio-temporal model to terrorism data in Iraq and demonstrate how choice of process model leads to differing conclusions on the existence of self-excitation in the data and differing conclusions on how violence spread spatially-temporally in that country from 2003–2010.

REFERENCES

- BAKER III, J. A. and HAMILTON, L. H. (2006). *The Iraq Study Group Report*. Vintage Books.
- BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236.
- BRAITHWAITE, A. and JOHNSON, S. D. (2015). The battle for Baghdad: Testing hypotheses about insurgency from risk heterogeneity, repeat victimization, and denial policing approaches. *Terrorism and Political Violence* **27** 112–132.
- CHUNG, F. R. K. (1997). *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics **92**. Amer. Math. Soc., Providence, RI. [MR1421568](#)
- CRESSIE, N. and WIKLE, C. K. (2012). *Statistics for Spatio-Temporal Data*. Wiley, New York.
- FEARON, J. D. (2007). Iraq’s civil war. *Foreign Aff.* **86** 2.
- GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–760.
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–534.
- GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2014). *Bayesian Data Analysis* **2**. Chapman & Hall/CRC, Boca Raton, FL.

Key words and phrases. Laplace approximations, nonseparable space–time, spatial-temporal dependence, Bayesian, terrorism.

- HOFFMAN, B. (2006). Insurgency and counterinsurgency in Iraq. *Studies in Conflict & Terrorism* **29** 103–121.
- KAISER, M. S. and CRESSIE, N. (1997). Modeling Poisson variables with positive spatial dependence. *Statist. Probab. Lett.* **35** 423–432.
- LAFREE, G. and DUGAN, L. (2007). Introducing the global terrorism database. *Terrorism and Political Violence* **19** 181–204.
- LAUB, P. J., TAIMRE, T. and POLLETT, P. K. (2015). Hawkes Processes. Preprint. Available at [arXiv:1507.02822](https://arxiv.org/abs/1507.02822).
- LEE, Y. and NELDER, J. A. (1996). Hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser. B* **58** 619–678. [MR1410182](#)
- LEWIS, E., MOHLER, G., BRANTINGHAM, P. J. and BERTOZZI, A. L. (2012). Self-exciting point process models of civilian deaths in Iraq. *Security Journal* **25** 244–264.
- LINKE, A. M., WITMER, F. D. and O’LOUGHLIN, J. (2012). Space–time granger analysis of the war in Iraq: A study of coalition and insurgent action–reaction. *Int. Interact.* **38** 402–425.
- MIDLARSKY, M. I., CRENSHAW, M. and YOSHIDA, F. (1980). Why violence spreads. *International Studies Quarterly* **24** 262–298.
- MOHLER, G. (2013). Modeling and estimation of multi-source clustering in crime and security data. *Ann. Appl. Stat.* **7** 1525–1539. [MR3127957](#)
- MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. [MR2816705](#)
- POLSON, N. G. and SCOTT, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Anal.* **7** 887–902. [MR3000018](#)
- PORTER, M. D., WHITE, G. (2012). Self-exciting hurdle models for terrorist activity. *Ann. Appl. Stat.* **6** 106–124.
- PYTHON, A., ILLIAN, J., JONES-TODD, C. and BLANGIARDO, M. (2016). A Bayesian Approach to Modelling Fine-Scale Spatial Dynamics of Non-State Terrorism: World Study, 2002–2013. Preprint. Available at [arXiv:1610.01215](https://arxiv.org/abs/1610.01215).
- 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.
- SHORT, M. B., D’ORSOGNA, M. R., PASOUR, V. B., TITA, G. E., BRANTINGHAM, P. J., BERTOZZI, A. L. and CHAYES, L. B. (2008). A statistical model of criminal behavior. *Math. Models Methods Appl. Sci.* **18** 1249–1267.
- 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.
- TENCH, S., FRY, H. and GILL, P. (2016). Spatio-temporal patterns of IED usage by the Provisional Irish Republican Army. *European J. Appl. Math.* **27** 377–402.
- U.S. ARMY AND MARINE CORPS (2006). Counterinsurgency, FM 3-24, MCWP 3-33.5.
- WEIDMANN, N. B. and WARD, M. D. (2010). Predicting conflict in space and time. *Journal of Conflict Resolution* **54** 883–901.
- WILLIAMS, P. (2009). *Criminals, Militias, and Insurgents: Organized Crime in Iraq*. Strategic Studies Institute, Carlisle, PA.
- ZHUKOV, Y. M. (2012). Roads and the diffusion of insurgent violence: The logistics of conflict in Russia’s North Caucasus. *Political Geography* **31** 144–156.

EMPIRICAL ASSESSMENT OF PROGRAMS TO PROMOTE COLLABORATION: A NETWORK MODEL APPROACH¹

BY KATHERINE R. MCLAUGHLIN AND JOSHUA D. EMBREE

Oregon State University and RAND Corporation

Collaboration networks are thought to be desirable to foster both individual and population productivity. Often programs are implemented to promote collaboration, for example, at academic institutions. However, few tools are available to assess the efficacy of these programs, and very few are data-driven. We carried out a survey at California State University, San Marcos during the 2012–2013 academic year to measure five types of collaboration ties among professors in five science departments at the university over time. During the time period of study, professors participated in NIH-sponsored curriculum development activities with members of other departments. It was hypothesized that participation in these activities would also foster overall collaboration between these departments.

This survey enables the exploration of several methodological and theoretical challenges in network research. In this paper we develop a statistical approach to assess the impact of programmatic interventions on collaboration using model-assisted social network analysis. We derive and implement a hierarchical Bayesian approach to modeling error-prone responses in surveys and examine the effect of an intervention on network structure. Based on this analysis we find an increase in educational collaboration over time after adjusting for the length of time each professor had to form collaborative ties at the university.

REFERENCES

- BORGATTI, S. P., CARLEY, K. M. and KRACKHARDT, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Soc. Netw.* **28** 124–136.
- BUTTS, C. T. (2003). Network inference, error, and informant (in)accuracy: A Bayesian approach. *Soc. Netw.* **25** 103–140.
- BUTTS, C. T., ACTON, R. M. and MARCUM, C. S. (2012). Interorganizational collaboration in the hurricane Katrina response. *J. Soc. Struct.* **13**.
- CROSS, J. E., DICKMANN, E., NEWMAN-GONCHAR, R. and FAGAN, J. M. (2009). Using mixed-method design and network analysis to measure development of interagency collaboration. *Am J. Eval.* **30** 310–29.
- EISENBERG, M. and SWANSON, N. (1996). Organizational network analysis as a tool for program evaluation. *Evaluation & the Health Professions* **19** 488–506.
- FREY, B. B., LOHMEIER, J. H., LEE, S. W. and TOLLEFSON, N. (2006). Measuring collaboration among grant partners. *Am J. Eval.* **27** 383–392.

Key words and phrases. Assessment, collaboration, dynamic networks, error-prone responses, grant evaluation, hierarchical Bayesian modeling, latent network, model-assisted analysis, recall bias, social network analysis, survey design.

- GAJDA, R. (2004). Utilizing collaboration theory to evaluate strategic alliances. *Am J. Eval.* **25** 65–77.
- GAJDA, R. and KOLIBA, C. (2007). Evaluating the imperative of intraorganizational collaboration: A school improvement perspective. *Am J. Eval.* **28** 26–44.
- HANDCOCK, M. S. and GILE, K. J. (2010). Modeling social networks from sampled data. *Ann. Appl. Stat.* **4** 5–25. [MR2758082](#)
- HANDCOCK, M. S., HUNTER, D. R., BUTTS, C. T., GOODREAU, S. M., KRIVITSKY, P. N., BENDER-DEMOLL, S. and MORRIS, M. (2016). statnet: Software tools for the statistical analysis of network data. The Statnet Project (<http://www.statnet.org>). R package version 2016.9.
- KRIVITSKY, P. N. and HANDCOCK, M. S. (2014). A separable model for dynamic networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 29–46. [MR3153932](#)
- LOADER, C. (2013). locfit: Local regression, likelihood and density estimation. R package version 1.5-9.1.
- MARCUM, C. S., BEVC, C. A. and BUTTS, C. T. (2012). Mechanisms of control in emergent interorganizational networks. *The Policy Studies Journal* **40** 516–546.
- MARSDEN, P. V. (1990). Network data and measurement. *Annu. Rev. Sociol.* **16** 435–463.
- MCCULLOUGH, J. (1992). Evaluating grants programs for basic scientific research: Difficult measurement problems, few techniques and slow progress. *Am J. Eval.* **13** 95–101.
- SNIJDERS, T. A. B. (1996). Stochastic actor-oriented models for network change. *J. Math. Sociol.* **21** 149–172.
- WANG, D. J., SHI, X., MCFARLAND, D. A. and LESKOVEC, J. (2012). Measurement error in network data: A re-classification. *Soc. Netw.* **34** 396–409.
- WOODLAND, R. H. and HUTTON, M. S. (2012). Evaluating organizational collaborations: Suggested entry points and strategies. *Am J. Eval.* **33** 366–383.

The Annals of Applied Statistics

Next Issues

Special section in memory of Stephen E. Fienberg (1942–2016), AOAS Editor-in-Chief 2013–2015

On Stephen E. Fienberg as a discussant and a friend DONALD B. RUBIN
Fingerprint science JOSEPH B. KADANE

REGULAR PAPERS

Estimating cellular pathways from an ensemble of heterogeneous data sources
ALEXANDER M. FRANKS, FLORIAN MARKOWETZ AND EDOARDO M. AIROLDI

A testing based approach to the discovery of differentially correlated variable sets
KELLY BODWIN, KAI ZHANG AND ANDREW B. NOBEL

Biomarker assessment and combination with differential covariate effects and an unknown gold standard, with an application
to Alzheimer's disease ZHEYU WANG AND XIAO-HUA ANDREW ZHOU

Robust dependence modeling for high-dimensional covariance matrices with financial applications
ZHE ZHU AND ROY E. WELSCH

Network-based feature screening with applications to genome data . . . MENGYUN WU, LIPING ZHU AND XINGDONG FENG

Statistical shape analysis of simplified neuronal trees ADAM DUNCAN, ERIC KLASSEN AND ANUJ SRIVASTAVA

Covariate matching methods for testing and quantifying wind turbine updates
YEI EUN SHIN, YU DING AND JIANHUA HUANG

Non-stationary modeling of tail dependence of two subjects' concentration
KSHITIJ SHARMA, VALÉRIE CHAVEZ-DEMOULIN AND PIERRE DILLENBOURG

A spatially-varying stochastic differential equation model for animal movement
JAMES C. RUSSELL, EPHRAIM M. HANKS, MURALI HARAN AND DAVID HUGHES

Torus principal component analysis with applications to RNA structure
BENJAMIN ELTZNER, STEPHAN HUCKEMANN AND KANTI V. MARDIA

TPRM: Tensor partition regression models with applications in imaging biomarker detection
HONGTU ZHU, MICHELLE F. MIRANDA AND JOSEPH G. IBRAHIM

Complex-valued time series modeling for improved activation detection in fMRI studies
DANIEL WRIGHT ADRIAN, RANJAN MAITRA AND DANIEL B. ROWE

Optimal multilevel matching using network flows: An application to a summer reading intervention
SAMUEL PIMENTEL, LINDSAY PAGE, MATTHEW LENARD AND LUKE J. KEELE

Topological data analysis of single-trial electroencephalographic signals
YUAN WANG, HERNANDO OMBAO AND MOO K. CHUNG

Joint significance tests for mediation effects of socioeconomic adversity on adiposity via epigenetics . . YEN-TSUNG HUANG

Adaptive-weight burden test for associations between quantitative traits and genotype data with complex correlations
XIAOWEI WU, TING GUAN, DAJIANG J. LIU, LUIS G. LEON NOVELO AND DIPANKAR BANDYOPADHYAY

Bayesian aggregation of average data: An application in drug development
SEBASTIAN WEBER, ANDREW GELMAN, DANIEL LEE, MICHAEL BETANCOURT, AKI VEHTARI
AND AMY RACINE-POON

BayCount: A Bayesian decomposition method for inferring tumor heterogeneity using RNA-Seq counts
FANGZHENG XIE, MINGYUAN ZHOU AND YANXUN XU

Exploring the conformational space for protein folding with sequential Monte Carlo
SAMUEL WONG, JUN S. LIU AND SAMUEL KOU

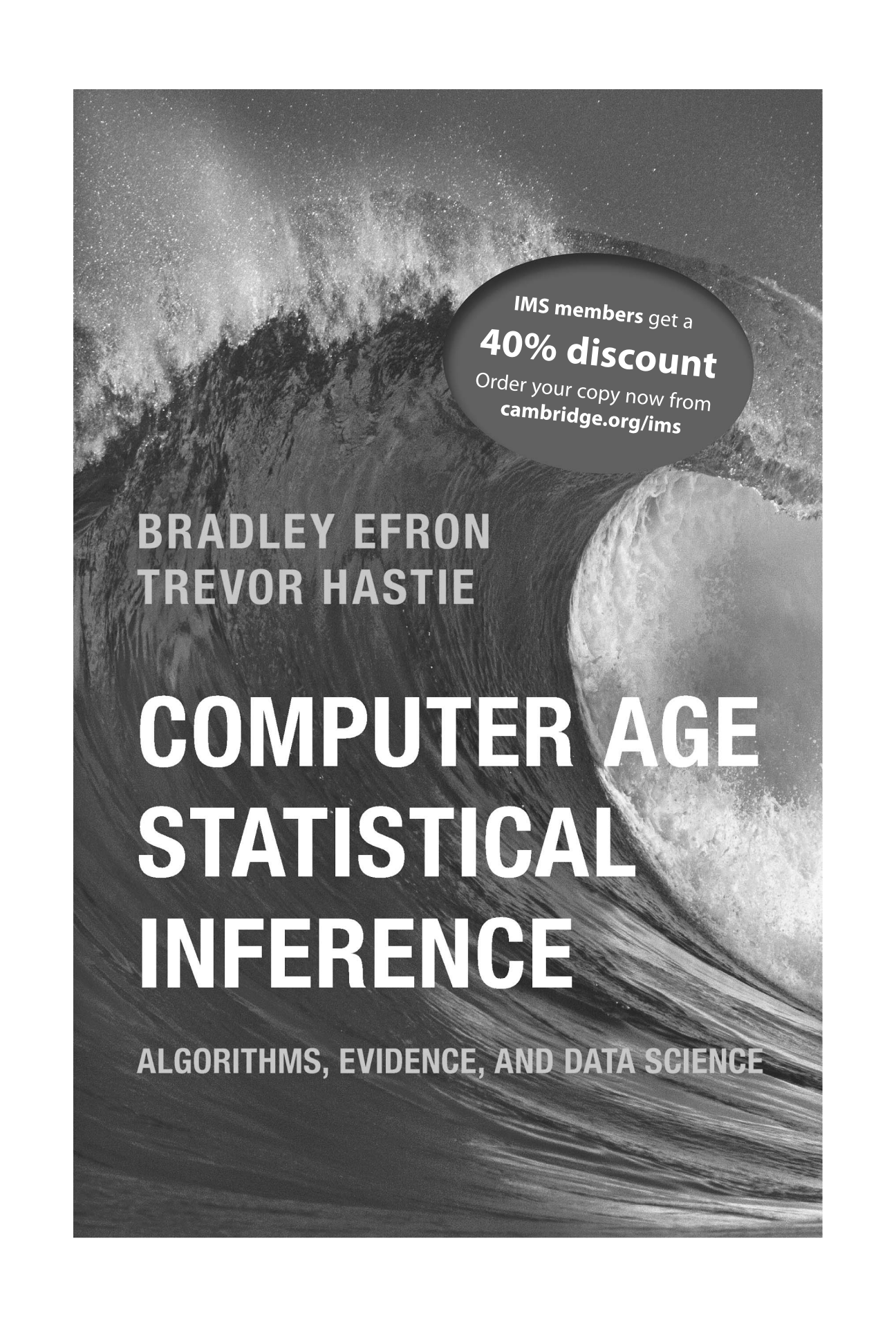
Sequential double cross-validation for assessment of added predictive ability in high-dimensional omic applications
MAR RODRIGUEZ-GIRONDO, PERTTU SALO, TOMASZ BURZYKOWSKI, MARKUS PEROLA,
JEANINE HOUWING-DUISTERMAAT AND BART MERTENS

Continued

The Annals of Applied Statistics

Next Issues—Continued

- Joining the incompatible: Exploiting purposive lists for the sample-based estimation of species richness
ALESSANDRO CHIARUCCI, ROSA MARIA DI BIASE, LORENZO FATTORINI, MARZIA MARCHESELLI
AND CATERINA PISANI
- A general framework for association analysis of heterogeneous data GEN LI AND IRINA GAYNANOVA
- Confident inference for SNP effects on treatment efficacy
YING DING, YING GRACE LI, YUSHI LIU, STEPHEN J. RUBERG AND JASON C. HSU
- Nonparametric Bayesian learning of heterogeneous dynamic transcription factor networks
XIANGYU LUO AND YINGYING WEI
- Estimating and comparing cancer progression risks under varying surveillance protocols
JANE M. LANGE, ROMAN GULATI, AMY S. LEONARDSON, DANIEL W. LIN, LISA F. NEWCOMB,
BRUCE J. TROCK, H. BALLENTINE CARTER, MATTHEW R. COOPERBERG, JANET E. COWAN,
LAWRENCE H. KLOTZ AND RUTH ETZIONI
- Analysing plant closure effects using time-varying mixture-of-experts Markov chain clustering
SYLVIA FRÜHWIRTH-SCHNATTER, STEFAN PITTNER, ANDREA WEBER AND RUDOLF WINTER-EBMER
- Using missing types to improve partial identification with application to a study of HIV prevalence in Malawi
ZHICHAO JIANG AND PENG DING
- A coupled ETAS-I2GMM point process with applications to seismic fault detection
YICHENG CHENG, MURAT DUNDAR AND GEORGE MOHLER
- Functional principal variance component testing for a genetic association study of HIV progression
DENIS AGNIEL, WEN XIE, MYRON ESSEX AND TIANXI CAI
- Estimating a common covariance matrix for network meta-analysis of gene expression datasets in diffuse large B-cell
lymphoma ANDERS ELLERN BILGRAU, RASMUS FROBERG BRØNDUM, PAUL SVANTE ERIKSEN,
KAREN DYBKÆR AND MARTIN BØGSTED
- Tree-based reinforcement learning for estimating optimal dynamic treatment regimes
YEBIN TAO, LU WANG AND DANIEL ALMIRALL
- A frequency-calibrated Bayesian search for new particles SHIRIN GOLCHI AND RICHARD A. LOCKHART
- Bayesian randomized response technique with multiple sensitive attributes: The case of information systems resource misuse
RAY S. W. CHUNG, AMANDA M. Y. CHU AND MIKE K. P. SO
- Direct likelihood-based inference for discretely observed stochastic compartmental models of infectious disease
LAM SI TUNG HO, FORREST CRAWFORD AND MARC SUCHARD
- Functional emulation of high resolution tsunami modelling over Cascadia
SERGE GUILLAS, ANDRIA SARRI, SIMON J. DAY, XIAOYU LIU AND FREDERIC DIAS
- Multi-rubric models for ordinal spatial data with application to online ratings data
ANTONIO RICARDO LINERO, JONATHAN R. BRADLEY AND APURVA DESAI
- The effects of non-ignorable missing data on label-free mass spectrometry proteomics experiments
JONATHON O'BRIEN, HARSHA GUNAWARDENA, JOAO PAULO, XIAN CHEN, JOSEPH IBRAHIM, STEVEN GYGI
AND BAHJAT QAQISH
- Bottom-up estimation and top-down prediction: Solar energy prediction combining information from multiple sources
YOUNGDEOK HWANG, SIYUAN LU AND JAE KWANG KIM



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