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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles

Clustering on the torus by conformal prediction
Sungkyu Jung, Kiho Park and Byungwon Kim 1583

Uncertainty quantification of a computer model for binary black hole formation
Luyao Lin, Derek Bingham, Floor Broekgaarden and Ilya Mandel 1604

Markov random field models for vector-based representations of landscapes
Patrizia Zamberletti, Julien Papaïx, Edith Gabriel and Thomas Opitz 1628

VCSEL: Prioritizing SNP-set by penalized variance component selection
Juhyun Kim, Judong Shen, Anran Wang, Devan V. Mehrotra, Seyoon Ko, Jin J. Zhou and Hua Zhou 1652

Zero-inflated quantile rank-score based test (ZIQRank) with application to scRNA-seq differential gene expression analysis . . . . . Wodan Ling, Wenfei Zhang, Bin Cheng and Ying Wei 1673

Joint and individual analysis of breast cancer histologic images and genomic covariates
Iain Carmichael, Benjamin C. Calhoun, Katherine A. Hoadley, Melissa A. Troester, Joseph Geradts, Heather D. Couture, Linnea Olsson, Charles M. Perou, Marc Niethammer, Jan Hannig and J. S. Marron 1697

Bayesian multistudy factor analysis for high-throughput biological data
Roberta De Vito, Ruggero Bellio, Lorenzo Trippa and Giovanni Parmigiani 1723

A Bayesian nonparametric approach to super-resolution single-molecule localization
Mariano I. Gabitto, Herve Marie-Nelly, Ari Pakman, Andras Pataki, Xavier Darzacq and Michael I. Jordan 1742

Bridging randomized controlled trials and single-arm trials using commensurate priors in arm-based network meta-analysis . . . . . Zhenxun Wang, Lifeng Lin, Thomas Murray, James S. Hodges and Haitao Chu 1767

Information content of high-order associations of the human gut microbiota network
Weston D. Viles, Juliette C. Madan, Hongzhe Li, Margaret R. Karagas and Anne G. Hoen 1788

RADIOHEAD: Radiogenomic analysis incorporating tumor heterogeneity in imaging through densities . . . . . Shariq Mohammed, Karthik Bharath, Sebastian Kurtek, Arvind Rao and Veerabhadran Baladandayuthapani 1808

Assessing the reliability of wind power operations under a changing climate with a non-Gaussian bias correction . . . . . Jiachen Zhang, Paola Crippa, Marc G. Genton and Stefano Castruccio 1831

Nonparametric importance sampling for wind turbine reliability analysis with stochastic computer models . . . . . Shuoran Li, Young Myoung Ko and Eunshin Byon 1850

Estimating animal utilization distributions from multiple data types: A joint spatiotemporal point process framework . . . . . Joe Watson, Ruth Joy, Dominic Tollit, Sheila J. Thornton and Marie Auger-Méthé 1872

continued

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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles—*Continued from front cover*

Spatial voting models in circular spaces: A case study of the U.S. House of Representatives . . . . . Xingchen Yu and Abel Rodríguez 1897
Modeling the social media relationships of Irish politicians using a generalized latent space stochastic blockmodel . . Tin Lok James Ng, Thomas Brendan Murphy, Ted Westling, Tyler H. McCormick and Bailey Fosdick 1923
Pan-disease clustering analysis of the trend of period prevalence . . . . . Sneha Jadhav, Chenjin Ma, Yefei Jiang, Ben-Chang Shia and Shuangge Ma 1945
Space-time smoothing models for subnational measles routine immunization coverage estimation with complex survey data . . . . . Tracy Qi Dong and Jon Wakefield 1959
Integrating geostatistical maps and infectious disease transmission models using adaptive multiple importance sampling . . . . . Renata Retkute, Panayiota Touloupou, María-Gloria Basáñez, T. Déirdre Hollingsworth and Simon E. F. Spencer 1980
Analysing the causal effect of London cycle superhighways on traffic congestion Prajamitra Bhuyan, Emma J. McCoy, Haojie Li and Daniel J. Graham 1999
The information in covariate imbalance in studies of hormone replacement therapy Ruoqi Yu, Dylan S. Small and Paul R. Rosenbaum 2023
Inferring food intake from multiple biomarkers using a latent variable model Silvia D'Angelo, Lorraine Brennan and Isobel Claire Gormley 2043
Mediation analysis for associations of categorical variables: The role of education in social class mobility in Britain . . . . . Jouni Kuha, Erzsébet Bukodi and John H. Goldthorpe 2061
Predicting competitions by combining conditional logistic regression and subjective Bayes: An Academy Awards case study Christopher T. Franck and Christopher E. Wilson 2083

THE ANNALS OF APPLIED STATISTICS Vol. 15, No. 4, pp. 1583–2100 December 2021

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: Krzysztof Burdzy, Department of Mathematics, University of Washington, Seattle, Washington 98195-4350, USA

President-Elect: Peter Bühlmann, Seminar für Statistik, ETH Zürich, 8092 Zürich, Switzerland

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

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

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

Program Secretary: Annie Qu, Department of Statistics, University of California, Irvine, Irvine, CA 92697-3425, USA

IMS PUBLICATIONS

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

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

The Annals of Probability. *Editors*: Alice Guionnet, Unité de Mathématiques Pures et Appliquées, ENS de Lyon, Lyon, France. Christophe Garban, Institut Camille Jordan, Université Claude Bernard Lyon 1, 69622 Villeurbanne, France

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

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

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

The Annals of Applied Statistics **[ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 15, Number 4, December 2021. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage pending at Cleveland, Ohio, and at additional mailing offices.

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

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1583–1603
https://doi.org/10.1214/21-AOAS1459

CLUSTERING ON THE TORUS BY CONFORMAL PREDICTION

BY SUNGKYU JUNG[1,*], KIHO PARK[1,†] AND BYUNGWON KIM[2]

[1]*Department of Statistics, Seoul National University,* [*]*sungkyu@snu.ac.kr;* [†]*pkh503201@snu.ac.kr*
[2]*Department of Statistics, Kyungpook National University, byungwonkim@knu.ac.kr*

Motivated by the analysis of torsion (dihedral) angles in the backbone of proteins, we investigate clustering of bivariate angular data on the torus $[-\pi, \pi) \times [-\pi, \pi)$. We show that naive adaptations of clustering methods, designed for vector-valued data, to the torus are not satisfactory and propose a novel clustering approach based on the conformal prediction framework. We construct several prediction sets for toroidal data with guaranteed finite-sample validity, based on a kernel density estimate and bivariate von Mises mixture models. From a prediction set built from a Gaussian approximation of the bivariate von Mises mixture, we propose a data-driven choice for the number of clusters and present algorithms for an automated cluster identification and cluster membership assignment. The proposed prediction sets and clustering approaches are applied to the torsion angles extracted from three strains of coronavirus spike glycoproteins (including SARS-CoV-2, contagious in humans). The analysis reveals a potential difference in the clusters of the SARS-CoV-2 torsion angles, compared to the clusters found in torsion angles from two different strains of coronavirus, contagious in animals.

REFERENCES

ARTHUR, D. and VASSILVITSKII, S. (2007). `k-means++`: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1027–1035. ACM, New York. MR2485254

BERG, J. M., TYMOCZKO, J. L. and STRYER, L. (2002). *Biochemistry*, 5th ed. W. H. Freeman & Company, New York.

BLUM, H. (1967). A transformation for extracting new descriptors of shape. In *Models for the Perception of Speech and Visual Form* (W. Wathen-Dunn, ed.) 362–380. MIT Press, Cambridge.

CHAKRABORTY, S. and WONG, S. W. (2017). BAMBI: An R package for fitting bivariate angular mixture models. arXiv preprint arXiv:1708.07804.

CHAN, J. F.-W., YUAN, S., KOK, K.-H., TO, K. K.-W., CHU, H., YANG, J., XING, F., LIU, J., YIP, C. C.-Y. et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. *Lancet* **395** 514–523.

CHENG, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **17** 790–799.

DI MARZIO, M., PANZERA, A. and TAYLOR, C. C. (2011). Kernel density estimation on the torus. *J. Statist. Plann. Inference* **141** 2156–2173. MR2772221 https://doi.org/10.1016/j.jspi.2011.01.002

DILL, K. A. and MACCALLUM, J. L. (2012). The protein-folding problem, 50 years on. *Science* **338** 1042–1046.

ELTZNER, B., HUCKEMANN, S. and MARDIA, K. V. (2018). Torus principal component analysis with applications to RNA structure. *Ann. Appl. Stat.* **12** 1332–1359. MR3834306 https://doi.org/10.1214/17-AOAS1115

GAO, Y., WANG, S., DENG, M. and XU, J. (2018). RaptorX-Angle: Real-value prediction of protein backbone dihedral angles through a hybrid method of clustering and deep learning. *BMC Bioinform.* **19** 100.

GONG, L., LI, J., ZHOU, Q., XU, Z., CHEN, L., ZHANG, Y., XUE, C., WEN, Z. and CAO, Y. (2017). A new bat-HKU2-like coronavirus in swine, China, 2017. *Emerg. Infec. Dis.* **23** 1607.

GORBALENYA, A. E., BAKER, S. C., BARIC, R. S. and CORONAVIRIDAE STUDY GROUP OF THE INTERNATIONAL COMMITTEE ON TAXONOMY OF VIRUSES (2020). The species severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **5** 536.

GRANT, B. J., RODRIGUES, A. P., ELSAWY, K. M., MCCAMMON, J. A. and CAVES, L. S. (2006). Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics* **22** 2695–2696.

Key words and phrases. Directional statistics, torsion angles, von Mises distribution, prediction set, density estimation, protein structure.

HARTIGAN, J. A. (1975). *Clustering Algorithms*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0405726

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

JUNG, S., PARK, K. and KIM, B. (2021). Supplement to "Clustering on the torus by conformal prediction." https://doi.org/10.1214/21-AOAS1459SUPPA, https://doi.org/10.1214/21-AOAS1459SUPPB

KAUFMAN, L. and ROUSSEEUW, P. J. (2009). *Finding Groups in Data*: *An Introduction to Cluster Analysis* **344**. John Wiley & Sons. MR1044997 https://doi.org/10.1002/9780470316801

KOUNTOURIS, P. and HIRST, J. D. (2009). Prediction of backbone dihedral angles and protein secondary structure using support vector machines. *BMC Bioinform.* **10** 437. https://doi.org/10.1186/1471-2105-10-437

LEI, J., RINALDO, A. and WASSERMAN, L. (2015). A conformal prediction approach to explore functional data. *Ann. Math. Artif. Intell.* **74** 29–43. MR3353895 https://doi.org/10.1007/s10472-013-9366-6

LEI, J., ROBINS, J. and WASSERMAN, L. (2013). Distribution-free prediction sets. *J. Amer. Statist. Assoc.* **108** 278–287. MR3174619 https://doi.org/10.1080/01621459.2012.751873

LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. and WASSERMAN, L. (2018). Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.* **113** 1094–1111. MR3862342 https://doi.org/10.1080/01621459.2017.1307116

LENNOX, K. P., DAHL, D. B., VANNUCCI, M. and TSAI, J. W. (2009). Density estimation for protein conformation angles using a bivariate von Mises distribution and Bayesian nonparametrics. *J. Amer. Statist. Assoc.* **104** 586–596. MR2751440 https://doi.org/10.1198/jasa.2009.0024

LOVELL, S. C., DAVIS, I. W., ARENDALL III, W. B., DE BAKKER, P. I., WORD, J. M., PRISANT, M. G., RICHARDSON, J. S. and RICHARDSON, D. C. (2003). Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins*: *Structure*, *Function*, *and Bioinformatics* **50** 437–450.

MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. Revised reprint of *Statistics of directional data* by Mardia [MR0336854 (49 #1627)]. MR1828667

MARDIA, K. V., TAYLOR, C. C. and SUBRAMANIAM, G. K. (2007). Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics* **63** 505–512. MR2370809 https://doi.org/10.1111/j.1541-0420.2006.00682.x

MARDIA, K. V., HUGHES, G., TAYLOR, C. C. and SINGH, H. (2008). A multivariate von Mises distribution with applications to bioinformatics. *Canad. J. Statist.* **36** 99–109. MR2432195 https://doi.org/10.1002/cjs.5550360110

MARDIA, K. V., KENT, J. T., ZHANG, Z., TAYLOR, C. C. and HAMELRYCK, T. (2012). Mixtures of concentrated multivariate sine distributions with applications to bioinformatics. *J. Appl. Stat.* **39** 2475–2492. MR2993298 https://doi.org/10.1080/02664763.2012.719221

MURTAGH, F. and CONTRERAS, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2** 86–97.

MURTAGH, F. and CONTRERAS, P. (2017). Algorithms for hierarchical clustering: An overview, II. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **7** e1219.

NODEHI, A., GOLALIZADEH, M., MAADOOLIAT, M. and AGOSTINELLI, C. (2021). Estimation of parameters in multivariate wrapped models for data on a p-torus. *Comput. Statist.* **36** 193–215. MR4215388 https://doi.org/10.1007/s00180-020-01006-x

NOURETDINOV, I., GAMMERMAN, J., FONTANA, M. and REHAL, D. (2020). Multi-level conformal clustering: A distribution-free technique for clustering and anomaly detection. *Neurocomputing* **397** 279–291.

O'NEILL, B. (2006). *Elementary Differential Geometry*, 2nd ed. Elsevier/Academic Press, Amsterdam. MR2351345

POLONIK, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Process. Appl.* **69** 1–24. MR1464172 https://doi.org/10.1016/S0304-4149(97)00028-8

SARGSYAN, K., WRIGHT, J. and LIM, C. (2012). GeoPCA: A new tool for multivariate analysis of dihedral angles based on principal component geodesics. *Nucleic Acids Res.* **40** e25–e25.

SCRUCCA, L., FOP, M., MURPHY, T. B. and RAFTERY, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8** 289–317.

SHAPOVALOV, M., VUCETIC, S. and DUNBRACK JR., R. L. (2019). A new clustering and nomenclature for beta turns derived from high-resolution protein structures. *PLoS Comput. Biol.* **15** e1006844.

SHIN, J., RINALDO, A. and WASSERMAN, L. (2019). Predictive clustering. arXiv preprint arXiv:1903.08125.

SINGH, H., HNIZDO, V. and DEMCHUK, E. (2002). Probabilistic model for two dependent circular variables. *Biometrika* **89** 719–723. MR1929175 https://doi.org/10.1093/biomet/89.3.719

VAN DER LAAN, M. J., POLLARD, K. S. and BRYAN, J. (2003). A new partitioning around medoids algorithm. *J. Stat. Comput. Simul.* **73** 575–584. MR1998670 https://doi.org/10.1080/0094965031000136012

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York. MR2161220

WALLS, A. C., PARK, Y.-J., TORTORICI, M. A., WALL, A., MCGUIRE, A. T. and VEESLER, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*.

WALTHER, D. and COHEN, F. E. (1999). Conformational attractors on the Ramachandran map. *Acta Crystallogr.*, *Sect. D*, *Biol. Crystallogr.* **55** 506–517.

XU, D. and TIAN, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science* **2** 165–193.

YU, J., QIAO, S., GUO, R. and WANG, X. (2020). Cryo-EM structures of HKU2 and SADS-CoV spike glycoproteins provide insights into coronavirus evolution. *Nat. Commun.* **11** 1–12.

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1604–1627
https://doi.org/10.1214/21-AOAS1484

UNCERTAINTY QUANTIFICATION OF A COMPUTER MODEL FOR BINARY BLACK HOLE FORMATION

BY LUYAO LIN[1,*], DEREK BINGHAM[1,†], FLOOR BROEKGAARDEN[2] AND ILYA MANDEL[3]

[1]*Department of Statistics and Actuarial Science, Simon Fraser University,* [*]*luyao_lin_2@sfu.ca;* [†]*derek_bingham@sfu.ca*
[2]*Harvard-Smithsonian Center for Astrophysics, floor.broekgaarden@cfa.harvard.edu*
[3]*School of Physics and Astronomy, Monash University, ilya.mandel@monash.edu*

In this paper, a fast and parallelizable method based on Gaussian processes (GPs) is introduced to emulate computer models that simulate the formation of binary black holes (BBHs) through the evolution of pairs of massive stars. Two obstacles that arise in this application are the a priori unknown conditions of BBH formation and the large scale of the simulation data. We address them by proposing a local emulator which combines a GP classifier and a GP regression model. The resulting emulator can also be utilized in planning future computer simulations through a proposed criterion for sequential design. By propagating uncertainties of simulation input through the emulator, we are able to obtain the distribution of BBH properties under the distribution of physical parameters.

REFERENCES

ANDREWS, J. J., ZEZAS, A. and FRAGOS, T. (2018). dart_board: Binary population synthesis with Markov chain Monte Carlo. *Astrophys. J., Suppl. Ser.* **237** 1. https://doi.org/10.3847/1538-4365/aaca30

BARRETT, J. W., MANDEL, I., NEIJSSEL, C. J., STEVENSON, S. and VIGNA-GÓMEZ, A. (2017). Exploring the parameter space of compact binary population synthesis. In *Astroinformatics* (M. Brescia, S. G. Djorgovski, E. D. Feigelson, G. Longo and S. Cavuoti, eds.). *IAU Symposium* **325** 46–50. https://doi.org/10.1017/S1743921317000059

BARRETT, J. W., GAEBEL, S. M., NEIJSSEL, C. J., VIGNA-GÓMEZ, A., STEVENSON, S., BERRY, C. P. L., FARR, W. M. and MANDEL, I. (2018). Accuracy of inference on the physics of binary evolution from gravitational-wave observations. *Mon. Not. R. Astron. Soc.* **477** 4685–4695. https://doi.org/10.1093/mnras/sty908

BELCZYNSKI, K., KALOGERA, V. and BULIK, T. (2002). A comprehensive study of binary compact objects as gravitational wave sources: Evolutionary channels, rates, and physical properties. *Astrophys. J.* **572** 407–431. https://doi.org/10.1086/340304

BINGHAM, D., RANJAN, P. and WELCH, W. J. (2014). Design of computer experiments for optimization, estimation of function contours, and related objectives. In *Statistics in Action* 109–124. CRC Press, Boca Raton, FL. MR3241971

BROEKGAARDEN, F. S., JUSTHAM, S., DE MINK, S. E., GAIR, J., MANDEL, I., STEVENSON, S., BARRETT, J. W., VIGNA-GÓMEZ, A. and NEIJSSEL, C. J. (2019). STROOPWAFEL: Simulating rare outcomes from astrophysical populations, with application to gravitational-wave sources. *Mon. Not. R. Astron. Soc.* **490** 5228–5248. https://doi.org/10.1093/mnras/stz2558

CONTI, S. and O'HAGAN, A. (2010). Bayesian emulation of complex multi-output and dynamic computer models. *J. Statist. Plann. Inference* **140** 640–651. MR2558393 https://doi.org/10.1016/j.jspi.2009.08.006

COVER, T. and HART, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13** 21–27.

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. MR2412639 https://doi.org/10.1111/j.1467-9868.2007.00633.x

DUNLOP, M. M., GIROLAMI, M. A., STUART, A. M. and TECKENTRUP, A. L. (2018). How deep are deep Gaussian processes? *J. Mach. Learn. Res.* **19** Paper No. 54, 46. MR3874162

FISHBACH, M. and HOLZ, D. E. (2017). Where are LIGO's big black holes? *Astrophys. J. Lett.* **851** L25.

GELBART, M. A., SNOEK, J. and ADAMS, R. P. (2014). Bayesian optimization with unknown constraints. arXiv preprint arXiv:1403.5607.

Key words and phrases. Computer experiments, surrogate model, local approximate GP, sequential design.

GRAMACY, R. B. and APLEY, D. W. (2015). Local Gaussian process approximation for large computer experiments. *J. Comput. Graph. Statist.* **24** 561–578. MR3357395 https://doi.org/10.1080/10618600.2014.914442

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

GRAMACY, R. B. and LEE, H. K. H. (2011). Optimization under unknown constraints. In *Bayesian Statistics* 9 229–256. Oxford Univ. Press, Oxford. With discussions by Christopher Holmes, M. Osborne, Antony Overstall, D. C. Woods and Daniel Williamson. MR3204008 https://doi.org/10.1093/acprof:oso/9780199694587.003.0008

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

HIGDON, D., GATTIKER, J., WILLIAMS, B. and RIGHTLEY, M. (2008). Computer model calibration using high-dimensional output. *J. Amer. Statist. Assoc.* **103** 570–583. MR2523994 https://doi.org/10.1198/016214507000000888

HSU, G. (2019). Fast emulation and calibration of large computer experiments with multivariate output. Unpublished M.Sc. Thesis, Dept. Statistics and Actuarial Science, Simon Fraser Univ.

JONES, D. R., SCHONLAU, M. and WELCH, W. J. (1998). Efficient global optimization of expensive black-box functions. *J. Global Optim.* **13** 455–492. MR1673460 https://doi.org/10.1023/A:1008306431147

KAUFMAN, C. G., BINGHAM, D., HABIB, S., HEITMANN, K. and FRIEMAN, J. A. (2011). Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology. *Ann. Appl. Stat.* **5** 2470–2492. MR2907123 https://doi.org/10.1214/11-AOAS489

KRUCKOW, M. U., TAURIS, T. M., LANGER, N., KRAMER, M. and IZZARD, R. G. (2018). Progenitors of gravitational wave mergers: Binary evolution with the stellar grid-based code COMBINE. *Mon. Not. R. Astron. Soc.* **481** 1908–1949. https://doi.org/10.1093/mnras/sty2190

LAWRENCE, E., HEITMANN, K., KWAN, J., UPADHYE, A., BINGHAM, D., HABIB, S., HIGDON, D., POPE, A., FINKEL, H. et al. (2017). The Mira–Titan universe. II. Matter power spectrum emulation. *Astrophys. J.* **847** 50.

LIN, L., BINGHAM, D., BROEKGAARDEN, F. and MANDEL, I. (2021). Supplement to "Uncertainty quantification of a computer model for binary black hole formation." https://doi.org/10.1214/21-AOAS1484SUPP

MANDEL, I. and FARMER, A. (2017). Gravitational waves: Stellar palaeontology. *Nature* **547** 284–285. https://doi.org/10.1038/547284a

MANDEL, I. and FARMER, A. (2018). Merging stellar-mass binary black holes. ArXiv E-prints.

MURRAY, I., PRESCOTT ADAMS, R. and MACKAY, D. J. (2010). Elliptical slice sampling.

NASH, J. E. and SUTCLIFFE, J. V. (1970). River flow forecasting through conceptual models part I—a discussion of principles. *J. Hydrol.* **10** 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

NEIJSSEL, C. J., VIGNA-GÓMEZ, A., STEVENSON, S., BARRETT, J. W., GAEBEL, S. M., BROEKGAARDEN, F. S., DE MINK, S. E., SZÉCSI, D., VINCIGUERRA, S. et al. (2019). The effect of the metallicity-specific star formation history on double compact object mergers. *Mon. Not. R. Astron. Soc.* **490** 3740–3759. https://doi.org/10.1093/mnras/stz2840

PETERS, P. C. and MATHEWS, J. (1963). Gravitational radiation from point masses in a Keplerian orbit. *Phys. Rev.* (2) **131** 435–440. MR0154716

QUIÑONERO-CANDELA, J. and RASMUSSEN, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.* **6** 1939–1959. MR2249877

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

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

STEVENSON, S., VIGNA-GÓMEZ, A., MANDEL, I., BARRETT, J. W., NEIJSSEL, C. J., PERKINS, D. and DE MINK, S. E. (2017). Formation of the first three gravitational-wave observations through isolated binary evolution. *Nat. Commun.* **8** 14906. https://doi.org/10.1038/ncomms14906

TANG, B. (1993). Orthogonal array-based Latin hypercubes. *J. Amer. Statist. Assoc.* **88** 1392–1397. MR1245375

TAYLOR, S. R. and GEROSA, D. (2018). Mining gravitational-wave catalogs to understand binary stellar evolution: A new hierarchical Bayesian framework. *Phys. Rev. D* **98** 083017. https://doi.org/10.1103/PhysRevD.98.083017

VIGNA-GÓMEZ, A., NEIJSSEL, C. J., STEVENSON, S., BARRETT, J. W., BELCZYNSKI, K., JUSTHAM, S., DE MINK, S. E., MÜLLER, B., PODSIADLOWSKI, P. et al. (2018). On the formation history of galactic double neutron stars. *Mon. Not. R. Astron. Soc.* **481** 4009–4029. https://doi.org/10.1093/mnras/sty2463

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1628–1651
https://doi.org/10.1214/21-AOAS1447

MARKOV RANDOM FIELD MODELS FOR VECTOR-BASED REPRESENTATIONS OF LANDSCAPES

BY PATRIZIA ZAMBERLETTI[*], JULIEN PAPAÏX[†], EDITH GABRIEL[‡] AND THOMAS OPITZ[§]

Biostatistique et Processus Spatiaux (BioSP) , INRAE, [*]patrizia.zamberletti@inrae.fr; [†]julien.papaix@inrae.fr; [‡]edith.gabriel@inrae.fr; [§]thomas.opitz@inrae.fr

In agricultural landscapes the spatial distribution of cultivated and semi-natural elements strongly impacts habitat connectivity and species dynamics. To allow for landscape structural analysis and scenario generation, we here develop statistical tools for real landscapes composed of geometric elements, including 2D patches but also 1D linear elements (e.g., hedges). Utilizing the framework of discrete Markov random fields, we design generative stochastic models that combine a multiplex network representation, based on spatial adjacency, with Gibbs energy terms to capture the distribution of landscape descriptors for land-use categories. We implement simulation of agricultural scenarios with parameter-controlled spatial and temporal patterns (e.g., geometry, connectivity, crop rotation), and we demonstrate through simulation that pseudo-likelihood estimation of parameters works well. To study statistical relevance of model components in real landscapes, we discuss model selection and validation, including cross-validated prediction scores. Model validation with a view toward ecologically relevant landscape summaries is achieved by comparing observed and simulated summaries (network metrics but also metrics and appropriately defined variograms using a raster discretization). Models fitted to subregions of the Lower Durance Valley (France) indicate strong deviation from random allocation and realistically capture landscape patterns. In summary, our approach improves the understanding of agroecosystems and enables simulation-based theoretical analysis of how landscape patterns shape biological and ecological processes.

REFERENCES

ADAMCZYK-CHAUVAT, K., KASSA, M., KIÊU, K., PAPAÏX, J. and STOICA, R. S. (2020). Gibbsian t-tessellation model for agricultural landscape characterization. Preprint. Available at arXiv:2007.16094.

BADDELEY, A. and MØLLER, J. (1989). Nearest-neighbour Markov point processes and random sets. *International Statistical Review/Revue Internationale de Statistique* 89–121.

BELGRANO, A., WOODWARD, G. and JACOB, U. (2015). *Aquatic Functional Biodiversity*: *An Ecological and Evolutionary Perspective*. Academic Press, San Diego, CA.

BESAG, J. E. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *J. Roy. Statist. Soc. Ser. B* **34** 75–83. MR0323276

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

BOCCALETTI, S., BIANCONI, G., CRIADO, R., DEL GENIO, C. I., GÓMEZ-GARDEÑES, J., ROMANCE, M., SENDIÑA-NADAL, I., WANG, Z. and ZANIN, M. (2014). The structure and dynamics of multilayer networks. *Phys. Rep.* **544** 1–122. MR3270140 https://doi.org/10.1016/j.physrep.2014.07.001

BONHOMME, V., CASTETS, M., IBANEZ, T., GÉRAUX, H., HÉLY, C. and GAUCHEREL, C. (2017). Configurational changes of patchy landscapes dynamics. *Ecol. Model.* **363** 1–7.

BOOTS, B., OKABE, A. and SUGIHARA, K. (1999). Spatial tessellations. *Geographical Information Systems* **1** 503–526.

BRIER, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78** 1–3.

Key words and phrases. Graphical model, Markov chain Monte Carlo simulation, multiplex-network, pseudo-likelihood, statistical landscape modeling, stochastic geometry.

BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **9**. IMS, Hayward, CA. MR0882001

BÜTTNER, G. and MAUCHA, G. (2006). *The Thematic Accuracy of CORINE Land Cover* 2000. *Assessment Using LUCAS* (*Land Use/Cover Area Frame Statistical Survey*). European Environment Agency, Copenhagen.

CALABRESE, J. M. and FAGAN, W. F. (2004). A comparison-shopper's guide to connectivity metrics. *Front. Ecol. Environ.* **2** 529–536.

CRESSIE, N. A. C. (1991). *Statistics for Spatial Data*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1127423

CRESSIE, N. A. C. (2015). *Statistics for Spatial Data*, Revised ed. *Wiley Classics Library*. Wiley, New York. MR3559472

CUSHMAN, S. A., MCGARIGAL, K. and NEEL, M. C. (2008). Parsimony in landscape metrics: Strength, universality, and consistency. *Ecol. Indicators* **8** 691–703.

CUSHMAN, S. A., GUTZWEILER, K., EVANS, J. S. and MCGARIGAL, K. (2010). The gradient paradigm: A conceptual and analytical framework for landscape ecology. In *Spatial Complexity*, *Informatics*, *and Wildlife Conservation* 83–108. Springer, New York.

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. *Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge Univ. Press, Cambridge. MR1478673 https://doi.org/10.1017/CBO9780511802843

ESTRADA, E. and BODIN, Ö. (2008). Using network centrality measures to manage landscape connectivity. *Ecol. Appl.* **18** 1810–1825.

FIENBERG, S. E. (2010). Introduction to papers on the modeling and analysis of network data. *Ann. Appl. Stat.* **4** 1–4. MR2758081 https://doi.org/10.1214/10-AOAS346

FORESIGHT, U. (2011). *The Future of Food and Farming*. *Final Project Report*. The Government Office for Science, London.

FRAZIER, A. E. and KEDRON, P. (2017). Landscape metrics: Past progress and future directions. *Current Landscape Ecology Reports* **2** 63–72.

GAETAN, C. and GUYON, X. (2010). *Spatial Statistics and Modeling*. *Springer Series in Statistics*. Springer, New York. MR2569034 https://doi.org/10.1007/978-0-387-92257-7

GALLAVOTTI, G. (1999). *Statistical Mechanics*: *A Short Treatise*. *Texts and Monographs in Physics*. Springer, Berlin. MR1707309 https://doi.org/10.1007/978-3-662-03952-6

GARDNER, R. H. (1999). RULE: Map generation and a spatial analysis program. In *Landscape Ecological Analysis* 280–303. Springer, New York.

GARDNER, R. H. and URBAN, D. L. (2007). Neutral models for testing landscape hypotheses. *Landsc. Ecol.* **22** 15–29.

GARDNER, R. H., MILNE, B. T., TURNEI, M. G. and O'NEILL, R. V. (1987). Neutral models for the analysis of broad-scale landscape pattern. *Landsc. Ecol.* **1** 19–28.

GARRIGUES, S., ALLARD, D., BARET, F. and WEISS, M. (2006). Quantifying spatial heterogeneity at the landscape scale using variogram models. *Remote Sens. Environ.* **103** 81–96.

GARRIGUES, S., ALLARD, D., BARET, F. and MORISETTE, J. (2008). Multivariate quantification of landscape spatial heterogeneity using variogram models. *Remote Sens. Environ.* **112** 216–230.

GAUCHEREL, C., FLEURY, D., AUCLAIR, D. and DREYFUS, P. (2006a). Neutral models for patchy landscapes. *Ecol. Model.* **197** 159–170.

GAUCHEREL, C., GIBOIRE, N., VIAUD, V., HOUET, T., BAUDRY, J. and BUREL, F. (2006b). A domain-specific language for patchy landscape modelling: The Brittany agricultural mosaic as a case study. *Ecol. Model.* **194** 233–243.

GAUCHEREL, C., BOUDON, F., HOUET, T., CASTETS, M. and GODIN, C. (2012). Understanding patchy landscape dynamics: Towards a landscape language. *PLoS ONE* **7** e46064. https://doi.org/10.1371/journal.pone.0046064

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

GREEN, P. J., HJORT, N. L. and RICHARDSON, S. (2003). *Highly Structured Stochastic Systems*, *Volume* 27. Oxford Univ. Press, Oxford.

GRELAUD, A., ROBERT, C. P., MARIN, J.-M., RODOLPHE, F. and TALY, J.-F. (2009). ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Anal.* **4** 317–335. MR2507366 https://doi.org/10.1214/09-BA412

HAMMERSLEY, J. M. and CLIFFORD, P. (1971). Markov fields on finite graphs and lattices. 46. Unpublished manuscript.

HESSELBARTH, M. H., SCIAINI, M., WITH, K. A., WIEGAND, K. and NOWOSAD, J. (2019). Landscapemetrics: An open-source R tool to calculate landscape metrics. *Ecography* **42** 1648–1657.

HIJMANS, R. J., VAN ETTEN, J., CHENG, J., MATTIUZZI, M., SUMNER, M., GREENBERG, J. A., LAMIGUEIRO, O. P., BEVAN, A., RACINE, E. B. et al. (2015). Package 'raster'. R package.
HOPCROFT, J. and TARJAN, R. (1973). Algorithm 447: Efficient algorithms for graph manipulation. *Commun. ACM* **16** 372–378.
INKOOM, J. N., FRANK, S., GREVE, K. and FÜRST, C. (2017). Designing neutral landscapes for data scarce regions in West Africa. *Ecol. Inform.* **42** 1–13.
JENSEN, J. L. and MØLLER, J. (1991). Pseudolikelihood for exponential family models of spatial point processes. *Ann. Appl. Probab.* **1** 445–461. MR1111528
KIÊU, K., ADAMCZYK-CHAUVAT, K., MONOD, H. and STOICA, R. S. (2013). A completely random T-tessellation model and Gibbsian extensions. *Spat. Stat.* **6** 118–138.
KIVELÄ, M., ARENAS, A., BARTHELEMY, M., GLEESON, J. P., MORENO, Y. and PORTER, M. A. (2014). Multilayer networks. *J. Complex Netw.* **2** 203–271.
KUPFER, J. A. (2012). Landscape ecology and biogeography: Rethinking landscape metrics in a post-FRAGSTATS landscape. *Progress in Physical Geography* **36** 400–420.
LAHIRI, S. N. (2003). *Resampling Methods for Dependent Data. Springer Series in Statistics*. Springer, New York. MR2001447 https://doi.org/10.1007/978-1-4757-3803-2
LANGHAMMER, M., THOBER, J., LANGE, M., FRANK, K. and GRIMM, V. (2019). Agricultural landscape generators for simulation models: A review of existing solutions and an outline of future directions. *Ecol. Model.* **393** 135–151.
LATORA, V. and MARCHIORI, M. (2001). Efficient behavior of small-world networks. *Phys. Rev. Lett.* **87** 198701.
LE BER, F., LAVIGNE, C., ADAMCZYK, K., ANGEVIN, F., COLBACH, N., MARI, J.-F. and MONOD, H. (2009). Neutral modelling of agricultural landscapes by tessellation methods—application for gene flow simulation. *Ecol. Model.* **220** 3536–3545.
LEFEBVRE, M., FRANCK, P., TOUBON, J.-F., BOUVIER, J.-C. and LAVIGNE, C. (2016). The impact of landscape composition on the occurrence of a canopy dwelling spider depends on orchard management. *Agriculture, Ecosystems & Environment* **215** 20–29.
LIN, Y., DENG, X., LI, X. and MA, E. (2014). Comparison of multinomial logistic regression and logistic regression: Which is more efficient in allocating land use? *Front. Earth Sci.* **8** 512–523.
LÜ, L., CHEN, D., REN, X.-L., ZHANG, Q.-M., ZHANG, Y.-C. and ZHOU, T. (2016). Vital nodes identification in complex networks. *Phys. Rep.* **650** 1–63. MR3543857 https://doi.org/10.1016/j.physrep.2016.06.007
MAALOULY, M., FRANCK, P., BOUVIER, J.-C., TOUBON, J.-F. and LAVIGNE, C. (2013). Codling moth parasitism is affected by semi-natural habitats and agricultural practices at orchard and landscape levels. *Agriculture, Ecosystems & Environment* **169** 33–42.
MARTIN, E. A., DAINESE, M., CLOUGH, Y., BÁLDI, A., BOMMARCO, R., GAGIC, V., GARRATT, M. P. D., HOLZSCHUH, A., KLEIJN, D. et al. (2019). The interplay of landscape composition and configuration: New pathways to manage functional biodiversity and agroecosystem services across Europe. *Ecol. Lett.* **22** 1083–1094. https://doi.org/10.1111/ele.13265
MCGARIGAL, K. and MARKS, B. J. (1995). FRAGSTATS: Spatial pattern analysis program for quantifying landscape structure. Gen. Tech. Rep. PNW-GTR-351. Portland, OR: US Department of Agriculture, Forest Service, Pacific Northwest Research Station. 122 p., 351.
MINOR, E. S. and URBAN, D. L. (2008). A graph-theory framework for evaluating landscape connectivity and conservation planning. *Conserv. Biol.* **22** 297–307.
MØLLER, J. and WAAGEPETERSEN, R. P. (1998). Markov connected component fields. *Adv. in Appl. Probab.* **30** 1–35. MR1618872 https://doi.org/10.1239/aap/1035227989
PAPAÏX, J., ADAMCZYK-CHAUVAT, K., BOUVIER, A., KIÊU, K., TOUZEAU, S., LANNOU, C. and MONOD, H. (2014). Pathogen population dynamics in agricultural landscapes: The ddal modelling framework. *Infect. Genet. Evol.* **27** 509–520.
POGGI, S., PAPAÏX, J., LAVIGNE, C., ANGEVIN, F., LE BER, F., PARISEY, N., RICCI, B., VINATIER, F. and WOHLFAHRT, J. (2018). Issues and challenges in landscape models for agriculture: From the representation of agroecosystems to the design of management strategies. *Landsc. Ecol.* **33** 1679–1690.
POWER, A. G. (2010). Ecosystem services and agriculture: Tradeoffs and synergies. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **365** 2959–2971.
RICCI, B., FRANCK, P., TOUBON, J.-F., BOUVIER, J.-C., SAUPHANOR, B. and LAVIGNE, C. (2009). The influence of landscape on insect pest dynamics: A case study in southeastern France. *Landsc. Ecol.* **24** 337–349.
SAURA, S. and MARTINEZ-MILLAN, J. (2000). Landscape patterns simulation with a modified random clusters method. *Landsc. Ecol.* **15** 661–678.
SCIAINI, M., FRITSCH, M., SCHERER, C. and SIMPKINS, C. E. (2018). NLMR and landscapetools: An integrated environment for simulating and modifying neutral landscape models in R. *Methods Ecol. Evol.* **9** 2240–2248.

SIRAMI, C., GROSS, N., BAILLOD, A. B., BERTRAND, C., CARRIÉ, R., HASS, A., HENCKEL, L., MIGUET, P., VUILLOT, C. et al. (2019). Increasing crop heterogeneity enhances multitrophic diversity across agricultural regions. *Proc. Natl. Acad. Sci. USA* **116** 16442–16447. https://doi.org/10.1073/pnas.1906419116

STOEHR, J. (2017). A review on statistical inference methods for discrete Markov random fields. Preprint. Available at arXiv:1704.03331.

URBAN, D. and KEITT, T. (2001). Landscape connectivity: A graph-theoretic perspective. *Ecology* **82** 1205–1218.

URBAN, D. L., MINOR, E. S., TREML, E. A. and SCHICK, R. S. (2009). Graph models of habitat mosaics. *Ecol. Lett.* **12** 260–273.

VAN LIESHOUT, M. N. M. (2000). *Markov Point Processes and Their Applications*. Imperial College Press, London. MR1789230 https://doi.org/10.1142/9781860949760

VAN LIESHOUT, M. N. M. (2019). *Theory of Spatial Statistics*: *A Concise Introduction*. CRC Press, Boca Raton, FL.

VARIN, C., REID, N. and FIRTH, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42. MR2796852

WITH, K. A. and KING, A. W. (1997). The use and misuse of neutral landscape models in ecology. *Oikos* **79** 219–229.

ZAMBERLETTI, P., PAPAÏX, J., GABRIEL, E. and OPITZ, T. (2021). Supplement to "Markov random field models for vector-based representations of landscapes." https://doi.org/10.1214/21-AOAS1447SUPP.

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1652–1672
https://doi.org/10.1214/21-AOAS1491

VCSEL: PRIORITIZING SNP-SET BY PENALIZED VARIANCE COMPONENT SELECTION

BY JUHYUN KIM[1,*], JUDONG SHEN[2,§], ANRAN WANG[2,¶], DEVAN V. MEHROTRA[2,‖], SEYOON KO[1,†], JIN J. ZHOU[3] AND HUA ZHOU[1,‡]

[1]*Department of Biostatistics, University of California, Los Angeles,* [*]*juhkim111@ucla.edu;* [†]*kos@ucla.edu;* [‡]*huazhou@ucla.edu*

[2]*Biostatistics and Research Decision Sciences, Merck & Co., Inc.,* [§]*judong.shen@merck.com;* [¶]*anran.wang@merck.com;* [‖]*devan_mehrotra@merck.com*

[3]*Department of Medicine, University of California, Los Angeles, jinjinzhou@ucla.edu*

Single nucleotide polymorphism (SNP) set analysis aggregates both common and rare variants and tests for association between phenotype(s) of interest and a set. However, multiple SNP-sets, such as genes, pathways, or sliding windows are usually investigated across the whole genome in which all groups are tested separately, followed by multiple testing adjustments. We propose a novel method to prioritize SNP-sets in a joint multivariate variance component model. Each SNP-set corresponds to a variance component (or kernel), and model selection is achieved by incorporating either convex or nonconvex penalties. The uniqueness of this variance component selection framework, which we call VCSEL, is that it naturally encompasses multivariate traits (VCSEL-M) and SNP-set-treatment or -environment interactions (VCSEL-I). We devise an optimization algorithm scalable to many variance components, based on the majorization-minimization (MM) principle. Simulation studies demonstrate the superiority of our methods in model selection performance, as measured by the area under the precision-recall (PR) curve, compared to the commonly used marginal testing and group penalization methods. Finally, we apply our methods to a real pharmacogenomics study and a real whole exome sequencing study. Some top ranked genes by VCSEL are detected as insignificant by the marginal test methods which emphasizes formal inference of individual genes with a strict significance threshold. This provides alternative insights for biologists to prioritize follow-up studies and develop polygenic risk score models.

REFERENCES

ABIFADEL, M., RABÈS, J.-P., DEVILLERS, M., MUNNICH, A., ERLICH, D., JUNIEN, C., VARRET, M. and BOILEAU, C. (2009). Mutations and polymorphisms in the proprotein convertase subtilisin kexin 9 (PCSK9) gene in cholesterol metabolism and disease. *Human Mutation* **30** 520–529.

BAKIN, S. (1999). Adaptive Regression and Model Selection in Data Mining Problems. Ph.D. thesis.

BANSAL, V., LIBIGER, O., TORKAMANI, A. and SCHORK, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* **11** 773–785. https://doi.org/10.1038/nrg2867

BATES, D. M. and PINHEIRO, J. C. (1998). Computational methods for multilevel modelling. *University of Wisconsin, Madison, WI* 1–29.

BENN, M., NORDESTGAARD, B. G., JENSEN, J. S., GRANDE, P., SILLESEN, H. and TYBJÆRG-HANSEN, A. (2005). Polymorphism in APOB associated with increased low-density lipoprotein levels in both genders in the general population. *The Journal of Clinical Endocrinology & Metabolism* **90** 5797–5803.

BEZANSON, J., EDELMAN, A., KARPINSKI, S. and SHAH, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Rev.* **59** 65–98. MR3605826 https://doi.org/10.1137/141000671

BODMER, W. and BONILLA, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40** 695–701. https://doi.org/10.1038/ng.f.136

Key words and phrases. Rare variants, majorization-minimization (MM), penalized estimation, variance components model, restricted maximum likelihood (REML), group selection, nonconvex penalties, multiple phenotypes.

BONDELL, H. D., KRISHNA, A. and GHOSH, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66** 1069–1077. MR2758494 https://doi.org/10.1111/j.1541-0420.2010.01391.x

BREHENY, P. and HUANG, J. (2009). Penalized methods for bi-level variable selection. *Stat. Interface* **2** 369–380. MR2540094 https://doi.org/10.4310/SII.2009.v2.n3.a10

BROADAWAY, K. A., DUNCAN, R., CONNEELY, K. N., ALMLI, L. M., BRADLEY, B., RESSLER, K. J. and EPSTEIN, M. P. (2015). Kernel approach for modeling interaction effects in genetic association studies of complex quantitative traits. *Genet. Epidemiol.* **39** 366–375.

BROADAWAY, K. A., CUTLER, D. J., DUNCAN, R., MOORE, J. L., WARE, E. B., JHUN, M. A., BIELAK, L. F., ZHAO, W., SMITH, J. A. et al. (2016). A statistical approach for testing cross-phenotype effects of rare variants. *Am. J. Hum. Genet.* **98** 525–540.

BUNESCU, R., GE, R., KATE, R. J., MARCOTTE, E. M., MOONEY, R. J., RAMANI, A. K. and WONG, Y. W. (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine* **33** 139–155.

CANNON, C. P., BLAZING, M. A., GIUGLIANO, R. P., MCCAGG, A., WHITE, J. A., THEROUX, P., DARIUS, H., LEWIS, B. S., OPHUIS, T. O. et al. (2015). Ezetimibe added to statin therapy after acute coronary syndromes. *N. Engl. J. Med.* **372** 2387–2397.

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

CHEN, Z. and DUNSON, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59** 762–769. MR2025100 https://doi.org/10.1111/j.0006-341X.2003.00089.x

CHEN, H., MEIGS, J. B. and DUPUIS, J. (2014). Incorporating gene-environment interaction in testing for association with rare genetic variants. *Hum. Hered.* **78** 81–90. https://doi.org/10.1159/000363347

CHOI, N. H., LI, W. and ZHU, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *J. Amer. Statist. Assoc.* **105** 354–364. With supplementary material available online. MR2656056 https://doi.org/10.1198/jasa.2010.tm08281

CINGOLANI, P., PLATTS, A., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. and RUDEN, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6** 80–92.

COHEN, J., PERTSEMLIDIS, A., KOTOWSKI, I. K., GRAHAM, R., GARCIA, C. K. and HOBBS, H. H. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* **37** 161–165.

CRAVEN, M. and BOCKHORST, J. (2005). Markov networks for detecting overalpping elements in sequence data. In *Advances in Neural Information Processing Systems* 193–200.

DAVIS, J. and GOADRICH, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* 233–240. ACM, New York.

DAVIS, J., BURNSIDE, E. S., DE CASTRO DUTRA, I., PAGE, D., RAMAKRISHNAN, R., COSTA, V. S. and SHAVLIK, J. W. (2005). View learning for statistical relational learning: With an application to mammography. In *IJCAI* 677–683. Citeseer.

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

DERING, C., HEMMELMANN, C., PUGH, E. and ZIEGLER, A. (2011). Statistical analysis of rare sequence variants: An overview of collapsing methods. *Genet. Epidemiol.* **35** S12–S17.

DESHMUKH, H. A., COLHOUN, H. M., JOHNSON, T., MCKEIGUE, P. M., BETTERIDGE, D. J., DURRINGTON, P. N., FULLER, J. H., LIVINGSTONE, S., CHARLTON-MENYS, V. et al. (2012). Genome-wide association study of genetic determinants of LDL-c response to atorvastatin therapy: Importance of Lp (a). *Journal of Lipid Research* **53** 1000–1011.

DUTTA, D., SCOTT, L., BOEHNKE, M. and LEE, S. (2019). Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes. *Genet. Epidemiol.* **43** 4–23.

FAN, Y. and LI, R. (2012). Variable selection in linear mixed effects models. *Ann. Statist.* **40** 2043–2068. MR3059076 https://doi.org/10.1214/12-AOS1028

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

GIBSON, G. (2012). Rare and common variants: Twenty arguments. *Nat. Rev. Genet.* **13** 135–145. https://doi.org/10.1038/nrg3118

GOADRICH, M., OLIPHANT, L. and SHAVLIK, J. (2004). Learning ensembles of first-order clauses for recall-precision curves: A case study in biomedical information extraction. In *International Conference on Inductive Logic Programming* 98–115. Springer, Berlin.

GUDMUNDSSON, J., SULEM, P., GUDBJARTSSON, D. F., MASSON, G., AGNARSSON, B. A., BENEDIKTSDOTTIR, K. R., SIGURDSSON, A., MAGNUSSON, O. T., GUDJONSSON, S. A. et al. (2012). A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat. Genet.* **44** 1326–1329.

HACKINGER, S. and ZEGGINI, E. (2017). Statistical methods to detect pleiotropy in human complex traits. *Open Biol.* **7**. https://doi.org/10.1098/rsob.170125

HARVILLE, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72** 320–340. With a comment by J. N. K. Rao and a reply by the author. MR0451550

HEID, I. M., BOES, E., MÜLLER, M., KOLLERITS, B., LAMINA, C., COASSIN, S., GIEGER, C., DÖRING, A., KLOPP, N. et al. (2008). Genome-wide association analysis of high-density lipoprotein cholesterol in the population-based KORA study sheds new light on intergenic regions. *Circ. Cardiovasc. Genet.* **1** 10–20.

HOFFMANN, T. J., THEUSCH, E., HALDAR, T., RANATUNGA, D. K., JORGENSON, E., MEDINA, M. W., KVALE, M. N., KWOK, P.-Y., SCHAEFER, C. et al. (2018). A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* **50** 401–413.

HOLMEN, O. L., ZHANG, H., FAN, Y., HOVELSON, D. H., SCHMIDT, E. M., ZHOU, W., GUO, Y., ZHANG, J., LANGHAMMER, A. et al. (2014). Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. *Nat. Genet.* **46** 345–351.

HUANG, J., MA, S., XIE, H. and ZHANG, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339–355. MR2507147 https://doi.org/10.1093/biomet/asp020

HUNTER, D. R. and LANGE, K. (2004). A tutorial on MM algorithms. *Amer. Statist.* **58** 30–37. MR2055509 https://doi.org/10.1198/0003130042836

KHURI, A. I. and SAHAI, H. (1985). Variance components analysis: A selective literature survey. *Int. Stat. Rev.* **53** 279–300. MR0967214 https://doi.org/10.2307/1402893

KIM, J., SHEN, J., WANG, A., MEHROTRA, D. V., KO, S., ZHOU, J. J. and ZHOU, H. (2021). Supplement to "VCSEL: Prioritizing SNP-set by penalized variance component selection." https://doi.org/10.1214/21-AOAS1491SUPPA, https://doi.org/10.1214/21-AOAS1491SUPPB.

KOK, S. and DOMINGOS, P. (2005). Learning the structure of Markov logic networks. In *Proceedings of the* 22*nd International Conference on Machine Learning* 441–448. ACM, New York.

LAIRD, N., LANGE, N. and STRAM, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *J. Amer. Statist. Assoc.* **82** 97–105. MR0883338

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 963–974.

LANGE, K. (2016). *MM Optimization Algorithms*. SIAM, Philadelphia, PA. MR3522165 https://doi.org/10.1137/1.9781611974409.ch1

LANGE, K., HUNTER, D. R. and YANG, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graph. Statist.* **9** 1–59. With discussion, and a rejoinder by Hunter and Lange. MR1819865 https://doi.org/10.2307/1390605

LANGE, L. A., HU, Y., ZHANG, H., XUE, C., SCHMIDT, E. M., TANG, Z.-Z., BIZON, C., LANGE, E. M., SMITH, J. D. et al. (2014). Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am. J. Hum. Genet.* **94** 233–245.

LEE, S., WU, M. C. and LIN, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13** 762–775.

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.

LEE, S., WON, S., KIM, Y. J., KIM, Y., CONSORTIUM, T.-G., KIM, B.-J. and PARK, T. (2017). Rare variant association test with multiple phenotypes. *Genet. Epidemiol.* **41** 198–209.

LI, B. and LEAL, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am. J. Hum. Genet.* **83** 311–321.

LIN, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika* **84** 309–326. MR1467049 https://doi.org/10.1093/biomet/84.2.309

LINDSTROM, M. J. and BATES, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Amer. Statist. Assoc.* **83** 1014–1022. MR0997577

MADSEN, B. E. and BROWNING, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5** e1000384. https://doi.org/10.1371/journal.pgen.1000384

MAITY, A., SULLIVAN, P. F. and TZENG, J.-I. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet. Epidemiol.* **36** 686–695.

MANNING, C. D. and SCHÜTZE, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA. MR1722790

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R. et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747.

PAILA, U., CHAPMAN, B. A., KIRCHNER, R. and QUINLAN, A. R. (2013). GEMINI: Integrative exploration of genetic variation and genome annotations. *PLoS Comput. Biol.* **9** e1003153. https://doi.org/10.1371/journal.pcbi.1003153

PATTERSON, H. D. and THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58** 545–554. MR0319325 https://doi.org/10.1093/biomet/58.3.545

PENG, H. and LU, Y. (2012). Model selection in linear mixed effect models. *J. Multivariate Anal.* **109** 109–129. MR2922858 https://doi.org/10.1016/j.jmva.2012.02.005

POSTMUS, I., TROMPET, S., DESHMUKH, H. A., BARNES, M. R., LI, X., WARREN, H. R., CHASMAN, D. I., ZHOU, K., ARSENAULT, B. J. et al. (2014). Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins. *Nat. Commun.* **5** 5068.

RAGHAVAN, V., BOLLMANN, P. and JUNG, G. S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems* (*TOIS*) **7** 205–229.

RIVAS, M. A. and MOUTSIANAS, L. (2015). Power of rare variant aggregate tests. In *Assessing Rare Variation in Complex Traits* 185–199. Springer, Berlin.

RIVAS, M. A., BEAUDOIN, M., GARDET, A., STEVENS, C., SHARMA, Y., ZHANG, C. K., BOUCHER, G., RIPKE, S., ELLINGHAUS, D. et al. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* **43** 1066–1073.

ROBINSON, D. L. (1987). Estimation and use of variance components. *Statistician* 3–14.

SAITO, T. and REHMSMEIER, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10** e0118432.

SCHAID, D. J., SINNWELL, J. P., LARSON, N. B. and CHEN, J. (2020). Penalized variance components for association of multiple genes with traits. *Genet. Epidemiol.* **44** 665–675. https://doi.org/10.1002/gepi.22340

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1190470 https://doi.org/10.1002/9780470316856

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. MR3173712 https://doi.org/10.1080/10618600.2012.681250

SINGLA, P. and DOMINGOS, P. (2005). Discriminative training of Markov logic networks. In *Proceedings of the* 20*th National Conference on Artificial Intelligene* (*AAAI*) **5** 868–873. AAAI Press, Menlo Park.

SIVAKUMARAN, S., AGAKOV, F., THEODORATOU, E., PRENDERGAST, J. G., ZGAGA, L., MANOLIO, T., RUDAN, I., MCKEIGUE, P., WILSON, J. F. et al. (2011). Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89** 607–618.

SOLOVIEFF, N., COTSAPAS, C., LEE, P. H., PURCELL, S. M. and SMOLLER, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nat. Rev. Genet.* **14** 483.

SOUTHAM, L., GILLY, A., SÜVEGES, D., FARMAKI, A.-E., SCHWARTZENTRUBER, J., TACHMAZIDOU, I., MATCHAN, A., RAYNER, N. W., TSAFANTAKIS, E. et al. (2017). Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat. Commun.* **8** 1–11.

SUDLOW, C., GALLACHER, J., ALLEN, N., BERAL, V., BURTON, P., DANESH, J., DOWNEY, P., ELLIOTT, P., GREEN, J. et al. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**.

SUO, C., TOULOPOULOU, T., BRAMON, E., WALSHE, M., PICCHIONI, M., MURRAY, R. and OTT, J. (2013). Analysis of multiple phenotypes in genome-wide genetic mapping studies. *BMC Bioinform.* **14** 151. https://doi.org/10.1186/1471-2105-14-151

SURAKKA, I., HORIKOSHI, M., MÄGI, R., SARIN, A.-P., MAHAJAN, A., LAGOU, V., MARULLO, L., FERREIRA, T., MIRAGLIO, B. et al. (2015). The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47** 589–597.

TACHMAZIDOU, I., DEDOUSSIS, G., SOUTHAM, L., FARMAKI, A.-E., RITCHIE, G. R., XIFARA, D. K., MATCHAN, A., HATZIKOTOULAS, K., RAYNER, N. W. et al. (2013). A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat. Commun.* **4** 1–6.

THOMPSON, W. A. JR. (1962). The problem of negative estimates of variance components. *Ann. Math. Stat.* **33** 273–289. MR0133912 https://doi.org/10.1214/aoms/1177704731

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

WALLACE, C., NEWHOUSE, S. J., BRAUND, P., ZHANG, F., TOBIN, M., FALCHI, M., AHMADI, K., DOBSON, R. J., MARÇANO, A. C. B. et al. (2008). Genome-wide association study identifies genes for biomarkers of cardiovascular disease: Serum urate and dyslipidemia. *Am. J. Hum. Genet.* **82** 139–149.

WOJCIK, G. L., GRAFF, M., NISHIMURA, K. K., TAO, R., HAESSLER, J., GIGNOUX, C. R., HIGHLAND, H. M., PATEL, Y. M., SOROKIN, E. P. et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570** 514–518.

WU, B. and PANKOW, J. S. (2016). Sequence kernel association test of multiple continuous phenotypes. *Genet. Epidemiol.* **40** 91–100. https://doi.org/10.1002/gepi.21945

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.

YANG, T., CHEN, H., TANG, H., LI, D. and WEI, P. (2019). A powerful and data-adaptive test for rare-variant-based gene-environment interaction analysis. *Stat. Med.* **38** 1230–1244. MR3920608 https://doi.org/10.1002/sim.8037

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 https://doi.org/10.1111/j.1467-9868.2005.00532.x

ZHAI, J., KIM, J., KNOX, K. S., TWIGG III, H. L., ZHOU, H. and ZHOU, J. J. (2018). Variance component selection with applications to microbiome taxonomic data. *Front. Microbiol.* **9** 509.

ZHAN, X., ZHAO, N., PLANTINGA, A., THORNTON, T. A., CONNEELY, K. N., EPSTEIN, M. P. and WU, M. C. (2017). Powerful genetic association analysis for common or rare variants with high-dimensional structured traits. *Genetics* **206** 1779–1790.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHANG, H., ZHAO, N., MEHROTRA, D. V. and SHEN, J. (2020). Composite kernel association test (CKAT) for SNP-set joint assessment of genotype and genotype-by-treatment interaction in pharmacogenetics studies. *Bioinformatics*. btaa125.

ZHAO, N., ZHANG, H., CLARK, J. J., MAITY, A. and WU, M. C. (2019). Composite kernel machine regression based on likelihood ratio test for joint testing of genetic and gene-environment interaction effect. *Biometrics* **75** 625–637. MR3999185 https://doi.org/10.1111/biom.13003

ZHOU, H., SEHL, M. E., SINSHEIMER, J. S. and LANGE, K. (2010). Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* **26** 2375.

ZHOU, H., HU, L., ZHOU, J. and LANGE, K. (2019). MM algorithms for variance components models. *J. Comput. Graph. Statist.* **28** 350–361. MR3974885 https://doi.org/10.1080/10618600.2018.1529601

ZUK, O., SCHAFFNER, S. F., SAMOCHA, K., DO, R., HECHTER, E., KATHIRESAN, S., DALY, M. J., NEALE, B. M., SUNYAEV, S. R. et al. (2014). Searching for missing heritability: Designing rare variant association studies. *Proc. Natl. Acad. Sci. USA* **111** E455–E464.

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1673–1696
https://doi.org/10.1214/21-AOAS1442

ZERO-INFLATED QUANTILE RANK-SCORE BASED TEST (ZIQRANK) WITH APPLICATION TO SCRNA-SEQ DIFFERENTIAL GENE EXPRESSION ANALYSIS

BY WODAN LING[1], WENFEI ZHANG[2], BIN CHENG[3,*] AND YING WEI[3,†]

[1]*Public Health Sciences Division, Fred Hutchinson Cancer Research Center, wling@fredhutch.org*
[2]*Sarepta Therapeutics, wenfeizhang@gmail.com*
[3]*Department of Biostatistics, Columbia University, *bc2159@cumc.columbia.edu; †yw2148@cumc.columbia.edu*

Differential gene expression analysis based on scRNA-seq data is challenging due to two unique characteristics of scRNA-seq data. First, multimodality and other heterogeneity of the gene expression among different cell conditions lead to divergences in the tail events or crossings of the expression distributions. Second, scRNA-seq data generally have a considerable fraction of dropout events, causing zero inflation in the expression. To account for the first characteristic, existing parametric approaches targeting the mean difference in gene expression are limited, while quantile regression that examines various locations in the distribution will improve the power. However, the second characteristic, zero inflation, makes the traditional quantile regression invalid and underpowered. We propose a quantile-based test that handles the two characteristics, multimodality and zero inflation, simultaneously. The proposed quantile rank-score based test for differential distribution detection (ZIQRank) is derived under a two-part quantile regression model for zero-inflated outcomes. It comprises a test in logistic modeling for the zero counts and a collection of rank-score tests adjusting for zero inflation at multiple prespecified quantiles of the positive part. The testing decision is based on an aggregate result by combining the marginal p-values by MinP or Cauchy procedure. The proposed test is asymptotically justified and evaluated with simulation studies. It shows a higher precision-recall AUC in detecting true differentially expressed genes (DEGs) than the existing methods. We apply the ZIQRank test to a TPM scRNA-seq data on human glioblastoma tumors and exclusively identify a group of DEGs between neoplastic and nonneoplastic cells, which are heterogeneous and have been proved to be associated with glioma. Application to a UMI count scRNA-seq data on cells from mouse intestinal organoids further demonstrates the capability of ZIQRank to improve and complement the existing approaches.

REFERENCES

BIRTWISTLE, M. R., RAUCH, J., KIYATKIN, A., AKSAMITIENE, E., DOBRZYŃSKI, M., HOEK, J. B., KOLCH, W., OGUNNAIKE, B. A. and KHOLODENKO, B. N. (2012). Emergence of bimodal cell population responses from the interplay between analog single-cell signaling and protein expression noise. *BMC Syst. Biol.* **6** 109. https://doi.org/10.1186/1752-0509-6-109

BUETTNER, F., NATARAJAN, K. N., CASALE, F. P., PROSERPIO, V., SCIALDONE, A., THEIS, F. J., TEICHMANN, S. A., MARIONI, J. C. and STEGLE, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33** 155–160. https://doi.org/10.1038/nbt.3102

COSTA-SILVA, J., DOMINGUES, D. and LOPES, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS ONE* **12** e0190152.

DARMANIS, S., SLOAN, S. A., CROOTE, D., MIGNARDI, M., CHERNIKOVA, S., SAMGHABABI, P., ZHANG, Y., NEFF, N., KOWARSKY, M. et al. (2017). Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep.* **21** 1399–1410.

Key words and phrases. Heterogeneity, multimodality, quantile rank-score based test, two-part model.

DOBRZYŃSKI, M., FEY, D., NGUYEN, L. K. and KHOLODENKO, B. N. (2012). Bimodal protein distributions in heterogeneous oscillating systems. In *International Conference on Computational Methods in Systems Biology* 17–28. Springer, Berlin.

DOBRZYŃSKI, M., NGUYEN, L. K., BIRTWISTLE, M. R., VON KRIEGSHEIM, A., FERNÁNDEZ, A. B., CHEONG, A., KOLCH, W. and KHOLODENKO, B. N. (2014). Nonlinear signalling networks and cell-to-cell variability transform external signals into broadly distributed or bimodal responses. *J. R. Soc. Interface* **11** 20140383. https://doi.org/10.1098/rsif.2014.0383

FAZI, B., FELSANI, A., GRASSI, L., MOLES, A., D'ANDREA, D., TOSCHI, N., SICARI, D., DE BONIS, P., ANILE, C. et al. (2015). The transcriptome and miRNome profiling of glioblastoma tissues and peritumoral regions highlights molecular pathways shared by tumors and surrounding areas and reveals differences between short-term and long-term survivors. *Oncotarget* **6** 22526.

FINAK, G., MCDAVID, A., YAJIMA, M., DENG, J., GERSUK, V., SHALEK, A. K., SLICHTER, C. K., MILLER, H. W., MCELRATH, M. J. et al. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16** 278.

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.

GUTENBRUNNER, C., JUREČKOVÁ, J., KOENKER, R. and PORTNOY, S. (1993). Tests of linear hypotheses based on regression rank scores. *J. Nonparametr. Stat.* **2** 307–331. MR1256383 https://doi.org/10.1080/10485259308832561

HE, Z., XU, B., LEE, S. and IONITA-LAZA, I. (2017). Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in metabochip data. *Am. J. Hum. Genet.* **101** 340–352.

HONG, S., CHEN, X., JIN, L. and XIONG, M. (2013). Canonical correlation analysis for RNA-seq co-expression networks. *Nucleic Acids Res.* **41** e95–e95.

ISLAM, S., KJÄLLQUIST, U., MOLINER, A., ZAJAC, P., FAN, J.-B., LÖNNERBERG, P. and LINNARSSON, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21** 1160–1167.

KÆRN, M., ELSTON, T. C., BLAKE, W. J. and COLLINS, J. J. (2005). Stochasticity in gene expression: From theories to phenotypes. *Nat. Rev. Genet.* **6** 451.

KANEHISA, M. and GOTO, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28** 27–30. https://doi.org/10.1093/nar/28.1.27

KHARCHENKO, P. V., SILBERSTEIN, L. and SCADDEN, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11** 740–742. https://doi.org/10.1038/nmeth.2967

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

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

KORTHAUER, K. D., CHU, L.-F., NEWTON, M. A., LI, Y., THOMSON, J., STEWART, R. and KENDZIORSKI, C. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* **17** 222.

LEE, S., WU, M. C. and LIN, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13** 762–775.

LING, W., ZHANG, W., CHENG, B. and WEI, Y. (2021). Supplement to "Zero-inflated quantile rank-score based test (ZIQRank) with application to scRNA-seq differential gene expression analysis." https://doi.org/10.1214/21-AOAS1442SUPPA, https://doi.org/10.1214/21-AOAS1442SUPPB

LIU, Y. and XIE, J. (2020). Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Amer. Statist. Assoc.* **115** 393–402. MR4078471 https://doi.org/10.1080/01621459.2018.1554485

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

MACHADO, J. A. F. and SANTOS SILVA, J. M. C. (2005). Quantiles for counts. *J. Amer. Statist. Assoc.* **100** 1226–1237. MR2236437 https://doi.org/10.1198/016214505000000330

MCKENZIE, L. D., LECLAIR, J. W., MILLER, K. N., STRONG, A. D., CHAN, H. L., OATES, E. L., LIGON, K. L., BRENNAN, C. W. and CHHEDA, M. G. (2019). CHD4 regulates the DNA damage response and RAD51 expression in glioblastoma. *Sci. Rep.* **9** 4444. https://doi.org/10.1038/s41598-019-40327-w

MOLIN, A. D., BARUZZO, G. and CAMILLO, B. D. (2017). Single-cell RNA-sequencing: Assessment of differential expression analysis methods. *Front. Genet.* **8** 62. https://doi.org/10.3389/fgene.2017.00062

MONK, N. A. (2003). Oscillatory expression of Hes1, p53, and NF-κB driven by transcriptional time delays. *Curr. Biol.* **13** 1409–1413.

OBACZ, J., ARCHAMBEAU, J., LE RESTE, P. J., PINEAU, R., JOUAN, F., BARROSO, K., VLACHAVAS, E., VOUTETAKIS, K., FAINSOD-LEVI, T. et al. (2019). IRE1-UBE2D3 signaling controls the recruitment of myeloid cells to glioblastoma. *BioRxiv* 533018.

PATEL, A. P., TIROSH, I., TROMBETTA, J. J., SHALEK, A. K., GILLESPIE, S. M., WAKIMOTO, H., CAHILL, D. P., NAHED, B. V., CURRY, W. T. et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344** 1396–1401.

RAMSKÖLD, D., LUO, S., WANG, Y.-C., LI, R., DENG, Q., FARIDANI, O. R., DANIELS, G. A., KHREBTUKOVA, I., LORING, J. F. et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30** 777.

SINGER, Z. S., YONG, J., TISCHLER, J., HACKETT, J. A., ALTINOK, A., SURANI, M. A., CAI, L. and ELOWITZ, M. B. (2014). Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Mol. Cell* **55** 319–331.

SONESON, C. and ROBINSON, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15** 255–261. https://doi.org/10.1038/nmeth.4612

SONG, X., LI, G., ZHOU, Z., WANG, X., IONITA-LAZA, I. and WEI, Y. (2017). QRank: A novel quantile regression tool for eQTL discovery. *Bioinformatics* **33** 2123–2130.

TENG, D.-C., SUN, J., AN, Y.-Q., HU, Z.-H., LIU, P., MA, Y.-C., HAN, B. and SHI, Y. (2016). Role of PHLPP1 in inflammation response: Its loss contributes to gliomas development and progression. *Int. Immunopharmacol.* **34** 229–234.

TOMBOLAN, L., POLI, E., MARTINI, P., ZIN, A., MILLINO, C., PACCHIONI, B., CELEGATO, B., BISOGNO, G., ROMUALDI, C. et al. (2016). Global DNA methylation profiling uncovers distinct methylation patterns of protocadherin alpha4 in metastatic and non-metastatic rhabdomyosarcoma. *BMC Cancer* **16** 886.

TRAPNELL, C., CACCHIARELLI, D., GRIMSBY, J., POKHAREL, P., LI, S., MORSE, M., LENNON, N. J., LIVAK, K. J., MIKKELSEN, T. S. et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32** 381.

TREUTLEIN, B., BROWNFIELD, D. G., WU, A. R., NEFF, N. F., MANTALAS, G. L., ESPINOZA, F. H., DESAI, T. J., KRASNOW, M. A. and QUAKE, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509** 371.

TROMBETTA, J. J., GENNERT, D., LU, D., SATIJA, R., SHALEK, A. K. and REGEV, A. (2014). Preparation of single-cell RNA-Seq libraries for next generation sequencing. *Curr. Protoc. Mol. Biol.* **107** 4–22.

UHLEN, M., OKSVOLD, P., FAGERBERG, L., LUNDBERG, E., JONASSON, K., FORSBERG, M., ZWAHLEN, M., KAMPF, C., WESTER, K. et al. (2010). Towards a knowledge-based human protein atlas. *Nat. Biotechnol.* **28** 1248.

WANG, H. J. (2009). Inference on quantile regression for heteroscedastic mixed models. *Statist. Sinica* **19** 1247–1261. MR2536154

WANG, H. and HE, X. (2007). Detecting differential expressions in GeneChip microarray studies: A quantile approach. *J. Amer. Statist. Assoc.* **102** 104–112. MR2293303 https://doi.org/10.1198/016214506000001220

WANG, T., LI, B., NELSON, C. E. and NABAVI, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinform.* **20** 40.

WEI, Y., PERE, A., KOENKER, R. and HE, X. (2006). Quantile regression methods for reference growth charts. *Stat. Med.* **25** 1369–1382. MR2226792 https://doi.org/10.1002/sim.2271

WEI, Y., SONG, X., LIU, M., IONITA-LAZA, I. and REIBMAN, J. (2016). Quantile regression in the secondary analysis of case-control data. *J. Amer. Statist. Assoc.* **111** 344–354. MR3494664 https://doi.org/10.1080/01621459.2015.1008101

ZHANG, Z. H., JHAVERI, D. J., MARSHALL, V. M., BAUER, D. C., EDSON, J., NARAYANAN, R. K., ROBINSON, G. J., LUNDBERG, A. E., BARTLETT, P. F. et al. (2014). A comparative study of techniques for differential expression analysis on RNA-Seq data. *PLoS ONE* **9** e103207.

ZHANG, W., WEI, Y., ZHANG, D. and XU, E. Y. (2020). ZIAQ: A quantile regression method for differential expression analysis of single-cell RNA-seq data. *Bioinformatics* **36** 3124–3130. https://doi.org/10.1093/bioinformatics/btaa098

ZHAO, Z. and XIAO, Z. (2014). Efficient regressions via optimally combining quantile information. *Econometric Theory* **30** 1272–1314. MR3278164 https://doi.org/10.1017/S0266466614000176

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1697–1722
https://doi.org/10.1214/20-AOAS1433

JOINT AND INDIVIDUAL ANALYSIS OF BREAST CANCER HISTOLOGIC IMAGES AND GENOMIC COVARIATES

BY IAIN CARMICHAEL[1], BENJAMIN C. CALHOUN[2], KATHERINE A. HOADLEY[3], MELISSA A. TROESTER[4], JOSEPH GERADTS[5], HEATHER D. COUTURE[6], LINNEA OLSSON[7], CHARLES M. PEROU[8], MARC NIETHAMMER[9], JAN HANNIG[10,*] AND J. S. MARRON[10,†]

[1]*Department of Statistics, University of Washington, idc9@uw.edu*

[2]*Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, ben.calhoun@unchealth.unc.edu*

[3]*Department of Genetics, Lineberger Comprehensive Cancer Center, Computational Medicine Program, University of North Carolina at Chapel Hill, hoadley@med.unc.edu*

[4]*Department of Epidemiology, Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, troester@unc.edu*

[5]*Department of Pathology and Laboratory Medicine, East Carolina University Brody School of Medicine, geradtsj20@ecu.edu*

[6]*Pixel Scientia Labs, heather@pixelscientia.com*

[7]*Department of Epidemiology, University of North Carolina at Chapel Hill, lolsson@live.unc.edu*

[8]*Department of Genetics, Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, cperou@med.unc.edu*

[9]*Department of Computer Science, University of North Carolina at Chapel Hill, mn@cs.unc.edu*

[10]*Department of Statistics, University of North Carolina at Chapel Hill, *jan.hannig@unc.edu; †marron@unc.edu*

The two main approaches in the study of breast cancer are histopathology (analyzing visual characteristics of tumors) and genomics. While both histopathology and genomics are fundamental to cancer research, the connections between these fields have been relatively superficial. We bridge this gap by investigating the Carolina Breast Cancer Study through the development of an integrative, exploratory analysis framework. Our analysis gives insights—some known, some novel—that are engaging to both pathologists and geneticists. Our analysis framework is based on angle-based joint and individual variation explained (AJIVE) for statistical data integration and exploits convolutional neural networks (CNNs) as a powerful, automatic method for image feature extraction. CNNs raise interpretability issues that we address by developing novel methods to explore visual modes of variation captured by statistical algorithms (e.g., PCA or AJIVE) applied to CNN features.

REFERENCES

ADEBAYO, J., GILMER, J., MUELLY, M., GOODFELLOW, I., HARDT, M. and KIM, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems* 9505–9515.

AEFFNER, F., ZARELLA, M. D., BUCHBINDER, N., BUI, M. M., GOODMAN, M. R., HARTMAN, D. J., LUJAN, G. M., MOLANI, M. A., PARWANI, A. V. et al. (2019). Introduction to digital image analysis in whole-slide imaging: A white paper from the digital pathology association. *J. Pathol. Inform.* **10**.

ALLOTT, E. H., GERADTS, J., COHEN, S. M., KHOURY, T., ZIRPOLI, G. R., BSHARA, W., DAVIS, W., OMILIAN, A., NAIR, P. et al. (2018). Frequency of breast cancer subtypes among African American women in the AMBER consortium. *Breast Cancer Res.* **20** 12.

ASH, J., DARNELL, G., MUNRO, D. and ENGELHARDT, B. (2018). Joint analysis of gene expression levels and histological images identifies genes associated with tissue morphology. *BioRxiv* 458711.

BACKENROTH, D., GOLDSMITH, J., HARRAN, M. D., CORTES, J. C., KRAKAUER, J. W. and KITAGO, T. (2018). Modeling motor learning using heteroscedastic functional principal components analysis. *J. Amer. Statist. Assoc.* **113** 1003–1015. MR3862335 https://doi.org/10.1080/01621459.2017.1379403

Key words and phrases. Multiview data, dimensionality reduction, image analysis, deep learning, interpretability, breast cancer histopathology, gene expression.

BECK, A. H., SANGOI, A. R., LEUNG, S., MARINELLI, R. J., NIELSEN, T. O., VAN DE VIJVER, M. J., WEST, R. B., VAN DE RIJN, M. and KOLLER, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.* **3** 108ra113. https://doi.org/10.1126/scitranslmed.3002564

BEJNORDI, B. E., MULLOOLY, M., PFEIFFER, R. M., FAN, S., VACEK, P. M., WEAVER, D. L., HERSCHORN, S., BRINTON, L. A., VAN GINNEKEN, B. et al. (2018). Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. *Mod. Pathol.* **31** 1502.

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

BISHOP, C. M. and TIPPING, M. E. (1998). A hierarchical latent variable model for data visualization. *IEEE Trans. Pattern Anal. Mach. Intell.* **20** 281–293.

CALDARELLA, A., BUZZONI, C., CROCETTI, E., BIANCHI, S., VEZZOSI, V., APICELLA, P., BIANCALANI, M., GIANNINI, A., URSO, C. et al. (2013). Invasive breast cancer: A significant correlation between histological types and molecular subgroups. *J. Cancer Res. Clin. Oncol.* **139** 617–623.

CAREY, L. A., PEROU, C. M., LIVASY, C. A., DRESSLER, L. G., COWAN, D., CONWAY, K., KARACA, G., TROESTER, M. A., TSE, C. K. et al. (2006). Race, breast cancer subtypes, and survival in the Carolina breast cancer study. *JAMA* **295** 2492–2502.

CARMICHAEL, I. (2020). pyjive: A Python package for AJIVE. Available at https://github.com/idc9/py_jive. https://doi.org/10.5281/zenodo.4091752

CARMICHAEL, I., CALHOUN, B. C., HOADLEY, K. A., TROESTER, M. A., GERADTS, J., COUTURE, H. D., OLSSON, L., PEROU, C. M., NIETHAMMER, M., HANNIG, J. and MARRON, J. S. (2021). Supplement to "Joint and individual analysis of breast cancer histologic images and genomic covariates." https://doi.org/10.1214/20-AOAS1433SUPPA, https://doi.org/10.1214/20-AOAS1433SUPPB, https://doi.org/10.1214/20-AOAS1433SUPPC.

CHEN, C., LI, O., TAO, C., BARNETT, A. J., SU, J. and RUDIN, C. (2018a). This looks like that: Deep learning for interpretable image recognition. Preprint. Available at arXiv:1806.10574.

CHEN, P.-H. C., GADEPALLI, K., MACDONALD, R., LIU, Y., NAGPAL, K., KOHLBERGER, T., DEAN, J., CORRADO, G. S., HIPP, J. D. et al. (2018b). Microscope 2.0: An augmented reality microscope with real-time artificial intelligence integration. Preprint. Available at arXiv:1812.00825.

CHEN, R. J., LU, M. Y., WANG, J., WILLIAMSON, D. F., RODIG, S. J., LINDEMAN, N. I. and MAHMOOD, F. (2019). Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. Preprint. Available at arXiv:1912.08937.

CHOLLET-HINTON, L., PUVANESARAJAH, S., SANDHU, R., KIRK, E. L., MIDKIFF, B. R., GHOSH, K., BRANDT, K. R., SCOTT, C. G., GIERACH, G. L. et al. (2018). Stroma modifies relationships between risk factor exposure and age-related epithelial involution in benign breast. *Mod. Pathol.* **31** 1085.

COLLEONI, M., ROTMENSZ, N., MAISONNEUVE, P., MASTROPASQUA, M. G., LUINI, A., VERONESI, P., INTRA, M., MONTAGNA, E., CANCELLO, G. et al. (2011). Outcome of special types of luminal breast cancer. *Ann. Oncol.* **23** 1428–1436.

COOPER, L. A. D., KONG, J., GUTMAN, D. A., DUNN, W. D., NALISNIK, M. and BRAT, D. J. (2015). Novel genotype-phenotype associations in human cancers enabled by advanced molecular platforms and computational analysis of whole slide images. *Lab. Invest.* **95** 366–376. https://doi.org/10.1038/labinvest.2014.153

COUDRAY, N., OCAMPO, P. S., SAKELLAROPOULOS, T., NARULA, N., SNUDERL, M., FENYÖ, D., MOREIRA, A. L., RAZAVIAN, N. and TSIRIGOS, A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24** 1559–1567. https://doi.org/10.1038/s41591-018-0177-5

COUTURE, H. D., WILLIAMS, L. A., GERADTS, J., NYANTE, S. J., BUTLER, E. N., MARRON, J. S., PEROU, C. M., TROESTER, M. A. and NIETHAMMER, M. (2018). Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *npj Breast Cancer* **4** 30.

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. and FEI-FEI, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 *IEEE Conference on Computer Vision and Pattern Recognition* 248–255. IEEE.

DI SAVERIO, S., GUTIERREZ, J. and AVISAR, E. (2008). A retrospective review with long term follow up of 11,400 cases of pure mucinous breast carcinoma. *Breast Cancer Res. Treat.* **111** 541–547.

DIAB, S. G., CLARK, G. M., OSBORNE, C. K., LIBBY, A., ALLRED, D. C. and ELLEDGE, R. M. (1999). Tumor characteristics and clinical outcome of tubular and mucinous breast carcinomas. *J. Clin. Oncol.* **17** 1442–1448. https://doi.org/10.1200/JCO.1999.17.5.1442

DRAPER, B., KIRBY, M., MARKS, J., MARRINAN, T. and PETERSON, C. (2014). A flag representation for finite collections of subspaces of mixed dimensions. *Linear Algebra Appl.* **451** 15–32. MR3198905 https://doi.org/10.1016/j.laa.2014.03.022

EIRO, N., GONZALEZ, L. O., FRAILE, M., CID, S., SCHNEIDER, J. and VIZOSO, F. J. (2019). Breast cancer tumor stroma: Cellular components, phenotypic heterogeneity, intercellular communication, prognostic implications and therapeutic opportunities. *Cancers* **11** 664.

ELMORE, J. G., LONGTON, G. M., CARNEY, P. A., GELLER, B. M., ONEGA, T., TOSTESON, A. N., NELSON, H. D., PEPE, M. S., ALLISON, K. H. et al. (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA* **313** 1122–1132.

ELSTON, C. W. and ELLIS, I. O. (2002). Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: Experience from a large study with long-term follow-up. CW Elston & IO Ellis. Histopathology 1991; 19; 403–410: AUTHOR COMMENTARY. *Histopathology* **41** 151–151.

FENG, Q., JIANG, M., HANNIG, J. and MARRON, J. S. (2018). Angle-based joint and individual variation explained. *J. Multivariate Anal.* **166** 241–265. MR3799646 https://doi.org/10.1016/j.jmva.2018.03.008

GAYNANOVA, I. and LI, G. (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics* **75** 1121–1132. MR4041816 https://doi.org/10.1111/biom.13108

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* 2672–2680.

HENG, Y. J., LESTER, S. C., TSE, G. M., FACTOR, R. E., ALLISON, K. H., COLLINS, L. C., CHEN, Y.-Y., JENSEN, K. C., JOHNSON, N. B. et al. (2017). The molecular basis of breast cancer pathological phenotypes. *J. Pathol.* **241** 375–391.

HOLZINGER, A., LANGS, G., DENK, H., ZATLOUKAL, K. and MÜLLER, H. (2019). Causability and explainabilty of artificial intelligence in medicine. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* e1312.

HOTELLING, H. (1936). Relation between two sets of variates. *Biometrika*.

HUNTER, J. D. (2007). Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9** 90.

ILSE, M., TOMCZAK, J. M. and WELLING, M. (2018). Attention-based deep multiple instance learning. Preprint. Available at arXiv:1802.04712.

JIMÉNEZ, G. and RACOCEANU, D. (2019). Deep learning for semantic segmentation versus classification in computational pathology: Application to mitosis analysis in breast cancer grading. *Front. Bioeng. Biotechnol.* **7** 145.

JOHNSTONE, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *Ann. Statist.* **36** 2638–2716. MR2485010 https://doi.org/10.1214/08-AOS605

JONES, E., OLIPHANT, T. and PETERSON, P. (2014). SciPy: Open source scientific tools for Python.

KETTENRING, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika* **58** 433–451. MR0341750 https://doi.org/10.1093/biomet/58.3.433

KIM, B., WATTENBERG, M., GILMER, J., CAI, C., WEXLER, J., VIEGAS, F. et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning* 2673–2682.

KINGMA, D. P. and WELLING, M. (2013). Auto-encoding variational Bayes. Preprint. Available at arXiv:1312.6114.

KOMURA, D. and ISHIKAWA, S. (2018). Machine learning methods for histopathological image analysis. *Comput. Struct. Biotechnol. J.* **16** 34–42. https://doi.org/10.1016/j.csbj.2018.01.001

LACROIX-TRIKI, M., SUAREZ, P. H., MACKAY, A., LAMBROS, M. B., NATRAJAN, R., SAVAGE, K., GEYER, F. C., WEIGELT, B., ASHWORTH, A. et al. (2010). Mucinous carcinoma of the breast is genomically distinct from invasive ductal carcinomas of no special type. *J. Pathol.* **222** 282–298.

LAZARD, D., SASTRE, X., FRID, M. G., GLUKHOVA, M. A., THIERY, J.-P. and KOTELIANSKY, V. E. (1993). Expression of smooth muscle-specific proteins in myoepithelium and stromal myofibroblasts of normal and malignant human breast tissue. *Proc. Natl. Acad. Sci. USA* **90** 999–1003.

LIU, Y., GADEPALLI, K., NOROUZI, M., DAHL, G. E., KOHLBERGER, T., BOYKO, A., VENUGOPALAN, S., TIMOFEEV, A., NELSON, P. Q. et al. (2017). Detecting cancer metastases on gigapixel pathology images. Preprint. Available at arXiv:1703.02442.

LIU, Y., KOHLBERGER, T., NOROUZI, M., DAHL, G. E., SMITH, J. L., MOHTASHAMIAN, A., OLSON, N., PENG, L. H., HIPP, J. D. et al. (2018). Artificial intelligence-based breast cancer nodal metastasis detection: Insights into the black box for pathologists. *Arch. Pathol. Lab. Med.*.

LIVASY, C. A., KARACA, G., NANDA, R., TRETIAKOVA, M. S., OLOPADE, O. I., MOORE, D. T. and PEROU, C. M. (2006). Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma. *Mod. Pathol.* **19** 264.

LOCK, E. F., HOADLEY, K. A., MARRON, J. S. and NOBEL, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7** 523–542. MR3086429 https://doi.org/10.1214/12-AOAS597

LU, M. Y., CHEN, R. J., WANG, J., DILLON, D. and MAHMOOD, F. (2019). Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. Preprint. Available at arXiv:1910.10825.

MACENKO, M., NIETHAMMER, M., MARRON, J. S., BORLAND, D., WOOSLEY, J. T., GUAN, X., SCHMITT, C. and THOMAS, N. E. (2009). A method for normalizing histology slides for quantitative analysis. In 2009 *IEEE International Symposium on Biomedical Imaging*: *From Nano to Macro* 1107–1110. IEEE.

MAHMOOD, F., YANG, Z., ASHLEY, T. and DURR, N. J. (2018). Multimodal densenet. Preprint. Available at arXiv:1811.07407.

MAHMOOD, F., BORDERS, D., CHEN, R., MCKAY, G. N., SALIMIAN, K. J., BARAS, A. and DURR, N. J. (2019). Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Trans. Med. Imag.*.

MCKINNEY, W. (2011). Pandas: A foundational Python library for data analysis and statistics. *Python High Perform. Sci. Comput.* **14**.

MOLNAR, C. et al. (2018). *Interpretable Machine Learning*: *A Guide for Making Black Box Models Explainable*. E-book at https://christophm.github.io/interpretable-ml-book/, version dated 10.

NETWORK, C. G. A. et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61.

OH, D. S., TROESTER, M. A., USARY, J., HU, Z., HE, X., FAN, C., WU, J., CAREY, L. A. and PEROU, C. M. (2006). Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. *J. Clin. Oncol.* **24** 1656–1664.

OLAH, C., SATYANARAYAN, A., JOHNSON, I., CARTER, S., SCHUBERT, L., YE, K. and MORDVINTSEV, A. (2018). The building blocks of interpretability. *Distill* **3** e10.

OORD, A. V. D., LI, Y. and VINYALS, O. (2018). Representation learning with contrastive predictive coding. Preprint. Available at arXiv:1807.03748.

OTSU, N. (1979). A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9** 62–66.

PARKER, J. S., MULLINS, M., CHEANG, M. C., LEUNG, S., VODUC, D., VICKERY, T., DAVIES, S., FAURON, C., HE, X. et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27** 1160.

PASZKE, A., GROSS, S., CHINTALA, S., CHANAN, G., YANG, E., DEVITO, Z., LIN, Z., DESMAISON, A., ANTIGA, L. et al. (2017). Automatic differentiation in pytorch.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A. et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12** 2825–2830. MR2854348

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

POURZANJANI, A. A., WU, T. B., JIANG, R. M., COHEN, M. J. and PETZOLD, L. R. (2017). Understanding coagulopathy using multi-view data in the presence of sub-cohorts: A hierarchical subspace approach. In *Machine Learning for Healthcare Conference* 338–351.

ROMÁN-PÉREZ, E., CASBAS-HERNÁNDEZ, P., PIRONE, J. R., REIN, J., CAREY, L. A., LUBET, R. A., MANI, S. A., AMOS, K. D. and TROESTER, M. A. (2012). Gene expression in extratumoral microenvironment predicts clinical outcome in breast cancer patients. *Breast Cancer Res.* **14** R51. https://doi.org/10.1186/bcr3152

ROSEN, P. P. (2001). *Rosen's Breast Pathology*. Williams & Wilkins, Baltimore.

SELVARAJU, R. R., COGSWELL, M., DAS, A., VEDANTAM, R., PARIKH, D. and BATRA, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In 2017 *IEEE International Conference on Computer Vision* (*ICCV*) 618–626. IEEE.

SHARIF RAZAVIAN, A., AZIZPOUR, H., SULLIVAN, J. and CARLSSON, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* 806–813.

SIMONYAN, K. and ZISSERMAN, A. (2014). Very deep convolutional networks for large-scale image recognition. Preprint. Available at arXiv:1409.1556.

SPRINGENBERG, J. T., DOSOVITSKIY, A., BROX, T. and RIEDMILLER, M. (2014). Striving for simplicity: The all convolutional net. Preprint. Available at arXiv:1412.6806.

SRIVASTAVA, A., KULKARNI, C., MALLICK, P., HUANG, K. and MACHIRAJU, R. (2018). Building trans-omics evidence: Using imaging and 'omics' to characterize cancer profiles. In *PSB* 377–388. World Scientific, Singapore.

SUNDARARAJAN, M., TALY, A. and YAN, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the* 34*th International Conference on Machine Learning*, *Vol.* 70 3319–3328. JMLR.org.

TROESTER, M. A., SUN, X., ALLOTT, E. H., GERADTS, J., COHEN, S. M., TSE, C.-K., KIRK, E. L., THORNE, L. B., MATHEWS, M. et al. (2017). Racial differences in PAM50 subtypes in the Carolina breast cancer study. *J. Natl. Cancer Inst.* **110** 176–182.

VAN DER WALT, S., COLBERT, S. C. and VAROQUAUX, G. (2011). The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* **13** 22.

VAN DER WALT, S., SCHÖNBERGER, J. L., NUNEZ-IGLESIAS, J., BOULOGNE, F., WARNER, J. D., YAGER, N., GOUILLART, E. and YU, T. (2014). scikit-image: Image processing in Python. *PeerJ* **2** e453. https://doi.org/10.7717/peerj.453

VELLIDO, A., MARTÍN-GUERRERO, J. D. and LISBOA, P. J. (2012). Making machine learning models interpretable. In *ESANN* **12** 163–172. Citeseer.

VETA, M., HENG, Y. J., STATHONIKOS, N., BEJNORDI, B. E., BECA, F., WOLLMANN, T., ROHR, K., SHAH, M. A., WANG, D. et al. (2019). Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. *Med. Image Anal.*.

WANG, C., PÉCOT, T., ZYNGER, D. L., MACHIRAJU, R., SHAPIRO, C. L. and HUANG, K. (2013). Identifying survival associated morphological features of triple negative breast cancer using multiple datasets. *J. Am. Med. Inform. Assoc.* **20** 680–687.

WANG, D., KHOSLA, A., GARGEYA, R., IRSHAD, H. and BECK, A. H. (2016). Deep learning for identifying metastatic breast cancer. Preprint. Available at arXiv:1606.05718.

WASKOM, M., BOTVINNIK, O., O'KANE, D., HOBSON, P., OSTBLOM, J., LUKAUSKAS, S., GEMPERLINE, D. C., AUGSPURGER, T., HALCHENKO, Y. et al. (2018). Seaborn (v0.9.0). https://doi.org/10.5281/zenodo.1313201

WEIGELT, B., GEYER, F. C., HORLINGS, H. M., KREIKE, B., HALFWERK, H. and REIS-FILHO, J. S. (2009). Mucinous and neuroendocrine breast carcinomas are transcriptionally distinct from invasive ductal carcinomas of no special type. *Mod. Pathol.* **22** 1401–1414. https://doi.org/10.1038/modpathol.2009.112

WEIN, L., SAVAS, P., LUEN, S. J., VIRASSAMY, B., SALGADO, R. and LOI, S. (2017). Clinical validity and utility of tumor-infiltrating lymphocytes in routine clinical practice for breast cancer patients: Current and future directions. *Front. Oncol.* **7** 156. https://doi.org/10.3389/fonc.2017.00156

WHITFIELD, M. L., SHERLOCK, G., SALDANHA, A. J., MURRAY, J. I., BALL, C. A., ALEXANDER, K. E., MATESE, J. C., PEROU, C. M., HURT, M. M. et al. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13** 1977–2000.

WILLIAMS, L. A., HOADLEY, K. A., NICHOLS, H. B., GERADTS, J., PEROU, C. M., LOVE, M. I., OLSHAN, A. F. and TROESTER, M. A. (2019). Differences in race, molecular and tumor characteristics among women diagnosed with invasive ductal and lobular breast carcinomas. *Cancer Causes Control* **30** 31–39. https://doi.org/10.1007/s10552-018-1121-1

WOLD, H. (1985). Partial least squares. In *Encyclopedia of Statistical Sciences*, *Vol.* 6 (S. Kotz and N. L. Johnson, eds.). Wiley, New York.

YANG, Z. and MICHAILIDIS, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* **32** 1–8. https://doi.org/10.1093/bioinformatics/btv544

YOSINSKI, J., CLUNE, J., BENGIO, Y. and LIPSON, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* 3320–3328.

ZACK, G. W., ROGERS, W. E. and LATT, S. A. (1977). Automatic measurement of sister chromatid exchange frequency. *J. Histochem. Cytochem.* **25** 741–753.

ZEILER, M. D. and FERGUS, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* 818–833. Springer, Berlin.

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1723–1741
https://doi.org/10.1214/21-AOAS1456

BAYESIAN MULTISTUDY FACTOR ANALYSIS FOR HIGH-THROUGHPUT BIOLOGICAL DATA

BY ROBERTA DE VITO[1], RUGGERO BELLIO[2], LORENZO TRIPPA[3] AND GIOVANNI PARMIGIANI[4]

[1]*Department of Biostatistics, Brown University, Roberta_DeVito@brown.edu*
[2]*Department of Economics and Statistics, University of Udine, ruggero.bellio@uniud.it*
[3]*Department of Data Science, Dana Farber Cancer Institute, ltrippa@jimmy.harvard.edu*
[4]*Department of Biostatistics, Harvard T. H. Chan School of Public Health, gp@jimmy.harvard.edu*

This paper analyzes breast cancer gene expression across seven studies to identify genuine and thus replicable gene patterns shared among these studies. Our premise is that genuine biological signal is more likely to be reproducibly present in multiple studies than spurious signal. Our analysis uses a new modeling strategy for the joint analysis of high-throughput biological studies which simultaneously identifies shared as well as study-specific signal. To this end, we generalize the multi-study factor analysis model to handle high-dimensional data and generalize the sparse Bayesian infinite factor model to this context. We provide strategies for the identification of the loading matrices, common and study-specific. Through extensive simulation analysis, we characterize the performance of the proposed approach in various scenarios and show that it outperforms standard factor analysis in identifying replicable signal in all scenarios considered. The analysis of breast cancer gene expression studies identifies clear replicable gene patterns. These patterns are related to well-known biological pathways involved in breast cancer, such as the ER, cell cycle, immune system, collagen, and metabolic pathways. Some of these patterns are also associated with existing breast cancer subtypes, such as LumA, Her2, and basal subtypes, while other patterns identify novel pathways active across subtypes and missed by hierarchical clustering approaches. The R package MSFA implementing the method is available on GitHub.

REFERENCES

AACH, J., RINDONE, W. and CHURCH, G. M. (2000). Systematic management and analysis of yeast gene expression data. *Genome Res*. **10** 431–445.

ABDI, H., WILLIAMS, L. J. and VALENTIN, D. (2013). Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. *Wiley Interdiscip. Rev.: Comput. Stat*. **5** 149–179.

ASSMANN, C., BOYSEN-HOGREFE, J. and PAPE, M. (2016). Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem. *J. Econometrics* **192** 190–206. MR3463672 https://doi.org/10.1016/j.jeconom.2015.10.010

BASSO, A. D., SOLIT, D. B., MUNSTER, P. N. and ROSEN, N. (2002). Ansamycin antibiotics inhibit Akt activation and cyclin D expression in breast cancer cells that overexpress HER2. *Oncogene* **21** 1159–1166.

BASTIAN, M., HEYMANN, S. and JACOMY, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Third International AAAI Conference on Weblogs and Social Media*.

BELIN, S. et al. (2009). Dysregulation of ribosome biogenesis and translational capacity is associated with tumor progression of human breast cancer cells. *PLoS ONE* **4** e7147.

BERNAU, C., RIESTER, M., BOULESTEIX, A.-L., PARMIGIANI, G., HUTTENHOWER, C., WALDRON, L. and TRIPPA, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* **30** i105–i112.

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

Key words and phrases. Dimension reduction, factor analysis, gene expression, Gibbs sampling, meta-analysis.

BLUM, Y., MIGNON, G. L., LAGARRIGUE, S. and CAUSEUR, D. (2010). A factor model to analyze heterogeneity in gene expression. *BMC Bioinform.* **11** 368. https://doi.org/10.1186/1471-2105-11-368
BUTTE, A. J., TAMAYO, P., SLONIM, D., GOLUB, T. R. and KOHANE, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA* **97** 12182–12186.
CARTER, C. L., ALLEN, C. and HENSON, D. E. (1989). Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer* **63** 181–187.
CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. MR2655722 https://doi.org/10.1198/016214508000000869
CHATTERJEE, G. et al. (2018). Molecular patterns of cancer colonisation in lymph nodes of breast cancer patients. *Breast Cancer Res.* **20** 143.
CIRIELLO, G., MILLER, M. L., AKSOY, B. A., SENBABAOGLU, Y., SCHULTZ, N. and SANDER, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45** 1127–1133.
DARTON, R. A. (1980). Rotation in factor analysis. *J. R. Stat. Soc., Ser. D, Stat.* **29** 167–194.
DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2019). Multi-study factor analysis. *Biometrics* **75** 337–346. MR3953734 https://doi.org/10.1111/biom.12974
DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2021). Supplement to "Bayesian multi-study factor analysis for high-throughput biological data." https://doi.org/10.1214/21-AOAS1456SUPPA, https://doi.org/10.1214/21-AOAS1456SUPPB
DE VITO, R. et al. (2019). Shared and study-specific dietary patterns and head and neck cancer risk in an international consortium. *Epidemiology* **30** 93–102.
DRAGHICI, S. et al. (2007). A systems biology approach for pathway level analysis. *Genome Res.* **17** 1537–1545.
DRAY, S., CHESSEL, D. and THIOULOUSE, J. (2003). Co-inertia analysis and the linking of ecological data tables. *Ecology* **84** 3078–3089.
DURANTE, D. (2017). A note on the multiplicative gamma process. *Statist. Probab. Lett.* **122** 198–204. MR3584158 https://doi.org/10.1016/j.spl.2016.11.014
EDEFONTI, V. et al. (2012). Nutrient-based dietary patterns and the risk of head and neck cancer: A pooled analysis in the International Head and Neck Cancer Epidemiology consortium. *Ann. Oncol.* **23** 1869–1880.
GAO, J., CIRIELLO, G., SANDER, C. and SCHULTZ, N. (2014). Collection, integration and analysis of cancer genomic profiles: From data to insight. *Curr. Option Genet. Dev.* **24** 92–98.
HAIBE-KAINS, B. et al. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *J. Natl. Cancer Inst.* **104** 311–325.
HAYES, D. N. et al. (2006). Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *J. Clin. Oncol.* **24** 5079–5090.
HICKS, S. C., TENG, M. and IRIZARRY, R. A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *BioRxiv*.
HUO, Z., DING, Y., LIU, S., OESTERREICH, S. and TSENG, G. (2016). Meta-analytic framework for sparse K-means to identify disease subtypes in multiple transcriptomic studies. *J. Amer. Statist. Assoc.* **111** 27–42. MR3494636 https://doi.org/10.1080/01621459.2015.1086354
HUTTENHOWER, C. et al. (2006). A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics* **22** 2890–2897.
IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264.
KAISER, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23** 187–200.
KERR, K. F. (2007). Extended analysis of benchmark datasets for agilent two-color microarrays. *BMC Bioinform.* **8** 371. https://doi.org/10.1186/1471-2105-8-371
KIM, S., KANG, D., HUO, Z., PARK, Y. and TSENG, G. C. (2017). Meta-analytic principal component analysis in integrative omics application. *Bioinformatics* **34** 1321–1328.
LARSEN, M. J., THOMASSEN, M., TAN, Q., SØRENSEN, K. P. and KRUSE, T. A. (2014). Microarray-based RNA profiling of breast cancer: Batch effect removal improves cross-platform consistency. *BioMed Research International* **2014**.
LAW, J. H. et al. (2008). Phosphorylated insulin-like growth factor-i/insulin receptor is present in all breast cancer subtypes and is related to poor survival. *Cancer Res.* **68** 10238–10246.
LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14** 41–67. MR2036762
MASUDA, H. et al. (2013). Differential response to neoadjuvant chemotherapy among 7 triple-negative breast cancer molecular subtypes. *Clin. Cancer Res.* **19** 5533–5540.

MENG, C., KUSTER, B., CULHANE, A. C. and GHOLAMI, A. M. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinform*. **15** 162. https://doi.org/10.1186/1471-2105-15-162

MOOTHA, V. K. et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet*. **34** 267–273.

NI, Y., MÜLLER, P., ZHU, Y. and JI, Y. (2018). Heterogeneous recipocal graphical models. *Biometrics* **74** 606–615. MR3825347 https://doi.org/10.1111/biom.12791

OOSHIMA, A., PARK, J. and KIM, S.-J. (2019). Phosphorylation status at Smad3 linker region modulates transforming growth factor-β-induced epithelial-mesenchymal transition and cancer progression. *Cancer Science* **110** 481.

PARK, T. (2005). A penalized likelihood approach to rotation of principal components. *J. Comput. Graph. Statist*. **14** 867–888. MR2211371 https://doi.org/10.1198/106186005X78134

PARKER, J. S. et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol*. **27** 1160–1167.

PEROU, C. M. et al. (2000). Molecular portraits of human breast tumours. *Nature* **406** 747.

PHAROAH, P. D. P. et al. (2013). GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat. Genet*. **45** 362–370.

PLANEY, C. R. and GEVAERT, O. (2016). CoINcIDE: A framework for discovery of patient subtypes across multiple datasets. *Gen. Med*. **8** 27. https://doi.org/10.1186/s13073-016-0281-4

REINERT, T., SAAD, E. D., BARRIOS, C. H. and BINES, J. (2017). Clinical implications of ESR1 mutations in hormone receptor-positive advanced breast cancer. *Front Oncol* **7** 26. https://doi.org/10.3389/fonc.2017.00026

RIESTER, M. et al. (2014). Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *J. Natl. Cancer Inst*. **106**.

ROBERT, P. and ESCOUFIER, Y. (1976). A unifying tool for linear multivariate statistical methods: The *RV*-coefficient. *J. R. Stat. Soc. Ser. C. Appl. Stat*. **25** 257–265. MR0440801 https://doi.org/10.2307/2347233

ROČKOVÁ, V. and GEORGE, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *J. Amer. Statist. Assoc*. **111** 1608–1622. MR3601721 https://doi.org/10.1080/01621459.2015.1100620

ROY, A., LAVINE, I., HERRING, A. H. and DUNSON, D. B. (2019). Perturbed factor analysis: Improving generalizability across studies. Preprint. Available at arXiv:1910.03021.

RUNCIE, D. E. and MUKHERJEE, S. (2013). Dissecting high-dimensional phenotypes with Bayesian sparse factor analysis of genetic covariance matrices. *Genetics* **194** 753–767.

SHI, L. et al. (2006). The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol*. **24** 1151–1161.

SØRLIE, T. et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* **98** 10869–10874.

SØRLIE, T. et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* **100** 8418–8423.

TYANOVA, S., ALBRECHTSEN, R., KRONQVIST, P., COX, J., MANN, M. and GEIGER, T. (2016). Proteomic maps of breast cancer subtypes. *Nat. Commun*. **7** 10259. https://doi.org/10.1038/ncomms10259

TYEKUCHEVA, S., MARCHIONNI, L., KARCHIN, R. and PARMIGIANI, G. (2011). Integrating diverse genomic data using gene sets. *Genome Biol*. **12** R105. https://doi.org/10.1186/gb-2011-12-10-r105

WANG, X. V., VERHAAK, R. G. W., PURDOM, E., SPELLMAN, P. T. and SPEED, T. P. (2011). Unifying gene expression measures from multiple platforms using factor analysis. *PLoS ONE* **6** e17691. https://doi.org/10.1371/journal.pone.0017691

YANG, F., LYU, S., DONG, S., LIU, Y., ZHANG, X. and WANG, O. (2016). Expression profile analysis of long noncoding RNA in HER-2-enriched subtype breast cancer by next-generation sequencing and bioinformatics. *OncoTargets and Therapy* **9** 761.

YOSHIDA, T. et al. (2015). CLK2 is an oncogenic kinase and splicing regulator in breast cancer. *Cancer Res*..

ZHANG, Y., BERNAU, C., PARMIGIANI, G. and WALDRON, L. (2020). The impact of different sources of heterogeneity on loss of accuracy from genomic prediction models. *Biostatistics* **21** 253–268. MR4133359 https://doi.org/10.1093/biostatistics/kxy044

ZHAO, S., GAO, C., MUKHERJEE, S. and ENGELHARDT, B. E. (2016). Bayesian group factor analysis with structured sparsity. *J. Mach. Learn. Res*. **17** Paper No. 196, 47. MR3580349

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1742–1766
https://doi.org/10.1214/21-AOAS1441

A BAYESIAN NONPARAMETRIC APPROACH TO SUPER-RESOLUTION SINGLE-MOLECULE LOCALIZATION

BY MARIANO I. GABITTO[1,*], HERVE MARIE-NELLY[2,‡], ARI PAKMAN[3], ANDRAS PATAKI[1,†], XAVIER DARZACQ[2,§] AND MICHAEL I. JORDAN[4]

[1]*Center for Computational Biology, Flatiron Institute, Simons Foundation,* [*]*mgabitto@simonsfoundation.org;* [†]*apataki@flatironinstitute.org*

[2]*Li Ka Shing Center for Biomedical and Health Sciences, University of California, Berkeley,* [‡]*hervemn@berkeley.edu;* [§]*darzacq@berkeley.edu*

[3]*Department of Statistics and Center for Theretical Neuroscience, Columbia University, aripakman@gmail.com*

[4]*Department of Statistics, University of California, Berkeley, jordan@stat.berkeley.edu*

We consider the problem of single-molecule identification in super-resolution microscopy. Super-resolution microscopy overcomes the diffraction limit by localizing individual fluorescing molecules in a field of view. This is particularly difficult since each individual molecule appears and disappears randomly across time and because the total number of molecules in the field of view is unknown. Additionally, data sets acquired with super-resolution microscopes can contain a large number of spurious fluorescent fluctuations caused by background noise.

To address these problems, we present a Bayesian nonparametric framework capable of identifying individual emitting molecules in super-resolved time series. We tackle the localization problem in the case in which each individual molecule is already localized in space. First, we collapse observations in time and develop a fast algorithm that builds upon the Dirichlet process. Next, we augment the model to account for the temporal aspect of fluorophore photophysics. Finally, we assess the performance of our methods with ground-truth data sets having known biological structure.

REFERENCES

ABBE, E. (1873). Beiträge zur theorie des mikroskops und der mikroskopischen wahrnehmung. *Arch. Mikrosk. Anat.* **9** 413–418.

AIRY, G. B. (1835). On the diffraction of an object-glass with circular aperture. *Trans. of the Cambridge Philosoph. Soc.* **5** 283–291.

ANNIBALE, P., SCARSELLI, M., KODIYAN, A. and RADENOVIC, A. (2010). Photoactivatable fluorescent protein mEos2 displays repeated photoactivation after a long-lived dark state in the red photoconverted form. *J. Phys. Chem. Lett.* **1** 1506–1510.

ANNIBALE, P., VANNI, S., SCARSELLI, M., ROTHLISBERGER, U. and RADENOVIC, A. (2011a). Identification of clustering artifacts in photoactivated localization microscopy. *Nat. Methods* **8** 527–528.

ANNIBALE, P., VANNI, S., SCARSELLI, M., ROTHLISBERGER, U. and RADENOVIC, A. (2011). Quantitative photo activated localization microscopy: Unraveling the effects of photoblinking. *PLoS ONE* **6** e22678. https://doi.org/10.1371/journal.pone.0022678

ARCHAMBEAU, C., LAKSHMINARAYANAN, B. and BOUCHARD, G. (2014). Latent IBP compound Dirichlet allocation. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 321–333.

BECK, M. and HURT, E. (2017). The nuclear pore complex: Understanding its function through structural insight. *Nat. Rev. Mol. Cel. Biol.* **18**.

BETZIG, E., PATTERSON, G. H., SOUGRAT, R., LINDWASSER, O. W., OLENYCH, S., BONIFACINO, J. S., DAVIDSON, M. W., LIPPINCOTT-SCHWARTZ, J. and HESS, H. F. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313** 1642–1645.

BLEI, D. M. and JORDAN, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1** 121–143. MR2227367 https://doi.org/10.1214/06-BA104

Key words and phrases. Super-resolution microscopy, Bayesian nonparametrics, variational inference.

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

BRAUN, M. and MCAULIFFE, J. (2010). Variational inference for large-scale models of discrete choice. *J. Amer. Statist. Assoc.* **105** 324–335. MR2757203 https://doi.org/10.1198/jasa.2009.tm08030

BRODERICK, T., JORDAN, M. I. and PITMAN, J. (2013). Cluster and feature modeling from combinatorial stochastic processes. *Statist. Sci.* **28** 289–312. MR3135534 https://doi.org/10.1214/13-STS434

D'ANGELO, M. A. and HETZER, M. W. (2008). Structure, dynamics and function of nuclear pore complexes. *Trends Cell Biol.* **18** 456–466.

DEMPSEY, G. T., VAUGHAN, J. C., CHEN, K. H., BATES, M. and ZHUANG, X. (2011). Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. *Nat. Methods* **8** 1027.

DESCHOUT, H. . (2014). Precisely and accurately localizing single emitters in fluorescence microscopy. *Nat. Methods* **11** 253–266.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

ESTER, M., KRIEGEL, H.-P., SANDER, J. and XU, X. (1996). A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. KDD'*96 226–231. AAAI Press, Menlo Park.

EWENS, W. J. (1990). Population genetics theory—the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory* (*Montreal*, *PQ*, 1987). *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* **299** 177–227. Kluwer Academic, Dordrecht. MR1108002

FAISAL, A., GILLBERG, J., PELTONEN, J., LEEN, G. and KASKI, S. (2012). Sparse nonparametric topic model for transfer learning. In *European Symposium on Artificial Neural Networks*.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

FINKEL, R. A. and BENTLEY, J. L. (1974). Quad trees a data structure for retrieval on composite keys. *Acta Inform.* **4** 1–9.

FOTI, N., FUTOMA, J., ROCKMORE, D. and WILLIAMSON, S. (2013). A unifying representation for a class of dependent random measures. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. *Proceedings of Machine Learning Research* **31**.

FOX, E., JORDAN, M. I., SUDDERTH, E. B. and WILLSKY, A. S. (2009). Sharing features among dynamical systems with beta processes. In *Advances in Neural Information Processing Systems* 549–557.

GABITTO, M. I., MARIE-NELLY, H., PAKMAN, A., PATAKI, A., DARZACQ, X. and JORDAN, M. I. (2021). Supplement to "A Bayesian nonparametric approach to super-resolution single-molecule localization." https://doi.org/10.1214/21-AOAS1441SUPPA, https://doi.org/10.1214/21-AOAS1441SUPPB

GAEL, J. V., TEH, Y. W. and GHAHRAMANI, Z. (2009). The infinite factorial hidden Markov model. In *Advances in Neural Information Processing Systems* 1697–1704.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

HANSEN, A. S., CATTOGLIO, C., DARZACQ, X. and TJIAN, R. (2018). Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus* **9** 20–32. https://doi.org/10.1080/19491034.2017.1389365

HEILEMANN, M., MARGEAT, E., KASPER, R., SAUER, M. and TINNEFELD, P. (2005). Carbocyanine dyes as efficient reversible single-molecule optical switch. *J. Am. Chem. Soc.* **127** 3801–3806.

HEILEMANN, M., DEDECKER, P., HOFKENS, J. and SAUER, M. (2009). Photoswitches: Key molecules for subdiffraction-resolution fluorescence imaging and molecular quantification. *Laser Photonics Rev.* **3** 180–202.

HOLDEN, S. J., UPHOFF, S. and KAPANIDIS, A. N. (2011). DAOSTORM: An algorithm for high-density super-resolution microscopy. *Nat. Methods* **8** 279–280. https://doi.org/10.1038/nmeth0411-279

HUANG, Z. L. . (2011). Localization-based super-resolution microscopy with an sCMOS camera. *Opt. Express* **19** 19156–19168.

HUGGINS, J. H. and WOOD, F. (2014). Infinite structured hidden semi-Markov models. Preprint. Available at arXiv:1407.0044.

HUGHES, M., KIM, D. I. and SUDDERTH, E. (2015). Reliable and scalable variational inference for the hierarchical Dirichlet process. In *Artificial Intelligence and Statistics* 370–378.

HUMMER, G., FRICKE, F. and HEILEMANN, M. (2016). Model-independent counting of molecules in single-molecule localization microscopy. *Mol. Biol. Cell* **27** 3637–3644. https://doi.org/10.1091/mbc.E16-07-0525

ICKSTADT, K. and WOLPERT, R. L. (1999). Spatial regression for marked point processes. In *Bayesian Statistics*, 6 (*Alcoceber*, 1998) 323–341. Oxford Univ. Press, New York. MR1723503

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

KAPLAN, C. and EWERS, H. (2015). Optimized sample preparation for single-molecule localization-based superresolution microscopy in yeast. *Nat. Protoc.* **10** 1007–1021.
KAPOOR-KAUSHIK, N. . (2016). Distinct mechanisms regulate Lck spatial organization in activated T cells. *Front. Immunol.* **7** 83.
KIM, S. J., FERNANDEZ-MARTINEZ, J., NUDELMAN, I., SHI, Y., ZHANG, W., RAVEH, B., HERRICKS, T., SLAUGHTER, B. D., HOGAN, J. A. et al. (2018). Integrative structure and functional anatomy of a nuclear pore complex. *Nature* **555** 475.
KOŠUTA, T., CULLELL-DALMAU, M., ZANACCHI, F. C. and MANZO, C. (2020). Bayesian analysis of data from segmented super-resolution images for quantifying protein clustering. *Phys. Chem. Chem. Phys.* **22** 1107–1114. https://doi.org/10.1039/c9cp05616e
KOTTAS, A. and SANSÓ, B. (2007). Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *J. Statist. Plann. Inference* **137** 3151–3163. MR2365118 https://doi.org/10.1016/j.jspi.2006.05.022
LEE, S.-H., SHIN, J. Y., LEE, A. and BUSTAMANTE, C. (2012). Counting single photoactivatable fluorescent molecules by photoactivated localization microscopy (PALM). *Proc. Natl. Acad. Sci. USA* **109** 17436–17441.
LI, Y., MUND, M., HOESS, P., DESCHAMPS, J., MATTI, U., NIJMEIJER, B., SABININA, V. J., ELLENBERG, J., SCHOEN, I. et al. (2018). Real-time 3D single-molecule localization using experimental point spread functions. *Nat. Methods* **15** 367–369. https://doi.org/10.1038/nmeth.4661
LIPPINCOTT-SCHWARTZ, J. and PATTERSON, G. H. (2009). Photoactivatable fluorescent proteins for diffraction-limited and super-resolution imaging. *Trends Cell Biol.* **19** 555–565.
NEAL, R. M. (1992). Bayesian mixture modeling. In *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis* **11** 197–211.
NEHME, E., FREEDMAN, D., GORDON, R., FERDMAN, B., WEISS, L. E., ALALOUF, O., ORANGE, R. and MICHAELI, T. (2019). DeepSTORM3D: Dense three dimensional localization microscopy and point spread function design by deep learning. *Int. J. Image Video Process.*
NICOVICH, P. R., OWEN, D. M. and GAUS, K. (2017). Turning single-molecule localization microscopy into a quantitative bioanalytical tool. *Nat. Protoc.* **12** 453–460. https://doi.org/10.1038/nprot.2016.166
NINO, D., RAFIEI, N., WANG, Y., ZILMAN, A. and MILSTEIN, J. N. (2017). Molecular counting with localization microscopy: A Bayesian estimate based on fluorophore statistics. *Biophys. J.* **112** 1777–1785.
OLIVIER, N., KELLER, D., GÖNCZY, P. and MANLEY, S. (2011). Resolution doubling in 3D-STORM imaging through improved buffers. *PLoS ONE* **8** e69004.
OVESNÝ, M., KŘÍŽEK, P., BORKOVEC, J., ŠVINDRYCH, Z. and HAGEN, G. M. (2014). ThunderSTORM: A comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics* **30** 2389–2390.
OWEN, D. M. . (2010). PALM imaging and cluster analysis of protein heterogeneity at the cell surface. *J. Biophotonics* **3** 446–454.
PERRONE, V., JENKINS, P. A., SPANÒ, D. and TEH, Y. W. (2017). Poisson random fields for dynamic feature models. *J. Mach. Learn. Res.* **18** Paper No. 127, 45. MR3763761
PERTSINIDIS, A., ZHANG, Y. and CHU, S. (2010). Subnanometre single-molecule localization, registration and distance measurements. *Nature* **466** 647–651.
PUCHNER, E. M., WALTER, J. M., KASPER, R., HUANG, B. and LIM, W. A. (2013). Counting molecules in single organelles with superresolution microscopy allows tracking of the endosome maturation trajectory. *Proc. Natl. Acad. Sci. USA* **110** 16015–16020.
RASMUSSEN, C. E. (2000). The infinite Gaussian mixture model. *Adv. Neural Inf. Process. Syst.* **12**.
REGIER, J., MILLER, A. C., SCHLEGEL, D., ADAMS, R. P., MCAULIFFE, J. D. and PRABHAT (2019). Approximate inference for constructing astronomical catalogs from images. *Ann. Appl. Stat.* **13** 1884–1926. MR4019161 https://doi.org/10.1214/19-AOAS1258
ROLLINS, G. C., SHIN, J. Y., BUSTAMANTE, C. and PRESSÉ, S. (2015). Stochastic approach to the molecular counting problem in superresolution microscopy. *Proc. Natl. Acad. Sci. USA* **112** E110–E118.
ROSSY, J., COHEN, E., GAUS, K. and OWEN, D. M. (2014). Method for co-cluster analysis in multichannel single-molecule localisation data. *Histochem. Cell Biol.* **141** 605–612.
ROSTEN, E., JONES, G. E. and COX, S. (2013). ImageJ plug-in for Bayesian analysis of blinking and bleaching. *Nat. Methods* **10** 97.
ROY, A., FIELD, M. J., ADAM, V. and BOURGEOIS, D. (2011). The nature of transient dark states in a photoactivatable fluorescent protein. *J. Am. Chem. Soc.* **133** 18586–18589.
RUBIN-DELANCHY, P., BURN, G. L., GRIFFIÉ, J., WILLIAMSON, D. J., HEARD, N. A., COPE, A. P. and OWEN, D. M. (2015). Bayesian cluster identification in single-molecule localization microscopy data. *Nat. Methods* **12** 1072–1076. https://doi.org/10.1038/nmeth.3612
RUST, M. J., BATES, M. and ZHUANG, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3** 793–795. https://doi.org/10.1038/nmeth929

SCHINDELIN, J., ARGANDA-CARRERAS, I., FRISE, E., KAYNIG, V., LONGAIR, M., PIETZSCH, T., PREIBISCH, S., RUEDEN, C., SAALFELD, S. et al. (2012). Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9** 676.

SCHUBERT, E., SANDER, J., ESTER, M., KRIEGEL, H.-P. and XU, X. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* **42** Art. 19, 21. MR3693646 https://doi.org/10.1145/3068335

SENGUPTA, P. . (2011). Probing protein heterogeneity in the plasma membrane using PALM and pair correlation analysis. *Nat. Methods* **8** 969–975.

SERGÉ, A., BERTAUX, N., RIGNEAULT, H. and MARGUET, D. (2008). Multiple-target tracing (MTT) algorithm probes molecular dynamics at cell surface. Protocol Exchange, https://doi.org/10.1038/nprot.2008.128.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SHCHERBAKOVA, D. M., SENGUPTA, P., LIPPINCOTT-SCHWARTZ, J. and VERKHUSHA, V. V. (2014). Photocontrollable fluorescent proteins for superresolution imaging. *Annu. Rev. Biophys.* **43** 303–329.

SMALL, A. and STAHLHEBER, S. (2014). Fluorophore localization algorithms for super-resolution microscopy. *Nat. Methods* **11** 267–279. https://doi.org/10.1038/nmeth.2844

SPECHT, C. G., IZEDDIN, I., RODRIGUEZ, P. C., EL BEHEIRY, M., ROSTAING, P., DARZACQ, X., DAHAN, M. and TRILLER, A. (2013). Quantitative nanoscopy of inhibitory synapses: Counting gephyrin molecules and receptor binding sites. *Neuron* **79** 308–321.

SPEISER, A., TURAGA, S. C. and MACKE, J. H. (2019). Teaching deep neural networks to localize sources in super-resolution microscopy by combining simulation-based learning and unsupervised learning. Available at arXiv:abs/1907.00770.

SUN, R., ARCHER, E. and PANINSKI, L. (2017). Scalable variational inference for super resolution microscopy. In *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics* 1057–1065.

SUN, S., PAISLEY, J. and LIU, Q. (2017). Location dependent Dirichlet processes. In *International Conference on Intelligent Science and Big Data Engineering* 64–76. Springer, Berlin.

SZYMBORSKA, A., DE MARCO, A., DAIGLE, N., CORDES, V. C., BRIGGS, J. A. and ELLENBERG, J. (2013). Nuclear pore scaffold structure analyzed by super-resolution microscopy and particle averaging. *Science* **341** 655–658.

TADDY, M. A. and KOTTAS, A. (2012). Mixture modeling for marked Poisson processes. *Bayesian Anal.* **7** 335–361. MR2934954 https://doi.org/10.1214/12-BA711

TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. MR2279480 https://doi.org/10.1198/016214506000000302

TURNER, R. E. and SAHANI, M. (2011). Two problems with variational expectation maximisation for time series models. In *Bayesian Time Series Models* 104–124. Cambridge Univ. Press, Cambridge. MR2894235

VALERA, I., RUIZ, F. J. and PEREZ-CRUZ, F. (2015). Infinite factorial unbounded-state hidden Markov model. *IEEE Trans. Pattern Anal. Mach. Intell.* **38** 1816–1828.

VAN DE LINDE, S., HEILEMANN, M. and SAUER, M. (2012). Live-cell super-resolution imaging with synthetic fluorophores. *Annu. Rev. Phys. Chem.* **63** 519–540.

VEATCH, S. L. . (2012). Correlation functions quantify super-resolution images and estimate apparent clustering due to over-counting. *PLoS ONE* **7** e31457.

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

WILLIAMSON, S., WANG, C., HELLER, K. A. and BLEI, D. M. (2010). The IBP compound Dirichlet process and its application to focused topic modeling. In *Proceedings of the* 27*th International Conference on Machine Learning* (*ICML*-10) 1151–1158. Citeseer.

XU, K., ZHONG, G. and ZHUANG, X. (2013). Actin, spectrin, and associated proteins form a periodic cytoskeletal structure in axons. *Science* **339** 452–456.

ZANACCHI, F. C., MANZO, C., ALVAREZ, A. S., DERR, N. D., GARCIA-PARAJO, M. F. and LAKADAMYALI, M. (2017). A DNA origami platform for quantifying protein copy number in super-resolution. *Nat. Methods* **14** 789–792. https://doi.org/10.1038/nmeth.4342

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1767–1787
https://doi.org/10.1214/21-AOAS1469

BRIDGING RANDOMIZED CONTROLLED TRIALS AND SINGLE-ARM TRIALS USING COMMENSURATE PRIORS IN ARM-BASED NETWORK META-ANALYSIS

BY ZHENXUN WANG[1,*], LIFENG LIN[2], THOMAS MURRAY[1,†], JAMES S. HODGES[1,‡] AND HAITAO CHU[1,§]

[1]*Division of Biostatistics, School of Public Health, University of Minnesota,* [*]*wang6795@umn.edu;* [†]*murra484@umn.edu;* [‡]*hodge003@umn.edu;* [§]*chux0051@umn.edu*

[2]*Department of Statistics, Florida State University, llin4@fsu.edu*

Network meta-analysis (NMA) is a powerful tool to compare multiple treatments directly and indirectly by combining and contrasting multiple independent clinical trials. Because many NMAs collect only a few eligible randomized controlled trials (RCTs), there is an urgent need to synthesize different sources of information, for example, from both RCTs and single-arm trials. However, single-arm trials and RCTs may have different populations and quality so that assuming they are exchangeable may be inappropriate. This article presents a novel method using a *commensurate prior on variance* (CPV) to borrow variance (rather than mean) information from single-arm trials in an arm-based (AB) Bayesian NMA. We illustrate the advantages of this CPV method by reanalyzing an NMA of immune checkpoint inhibitors in cancer patients. Comprehensive simulations investigate the impact on statistical inference of including single-arm trials. The simulation results show that the CPV method provides efficient and robust estimation, even when the two sources of information are moderately inconsistent.

REFERENCES

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

BEGG, C. B. and PILOTE, L. (1991). A model for incorporating historical controls into a meta-analysis. *Biometrics* **47** 899–906. MR1141952 https://doi.org/10.2307/2532647

CHEN, M.-H., IBRAHIM, J. G., LAM, P., YU, A. and ZHANG, Y. (2011). Bayesian design of noninferiority trials for medical devices using historical data. *Biometrics* **67** 1163–1170. MR2829252 https://doi.org/10.1111/j.1541-0420.2011.01561.x

DE VALPINE, P. (2016). Comparisons between NIMBLE, JAGS and Stan for the election88 example ("full" version) from the book known as "Applied Regression Modeling" (Gelman and Hill 2007). Accessed: 2020-12-18.

DE VALPINE, P., TUREK, D., PACIOREK, C. J., ANDERSON-BERGMAN, C., TEMPLE LANG, D. and BODIK, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *J. Comput. Graph. Statist.* **26** 403–413. MR3640196 https://doi.org/10.1080/10618600.2016.1172487

DIAS, S. and ADES, A. E. (2016). Absolute or relative effects? Arm-based synthesis of trial data. *Res. Synth. Methods* **7** 23–28.

DIAS, S., SUTTON, A. J., ADES, A. E. and WELTON, N. J. (2013). Evidence synthesis for decision making 2: A generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Med. Decis. Mak.* **33** 607–617.

DUAN, Y., YE, K. and SMITH, E. P. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics* **17** 95–106. MR2222036 https://doi.org/10.1002/env.752

EFTHIMIOU, O., MAVRIDIS, D., DEBRAY, T. P. A., SAMARA, M., BELGER, M., SIONTIS, G. C. M., LEUCHT, S. and SALANTI, G. (2017). Combining randomized and non-randomized evidence in network meta-analysis. *Stat. Med.* **36** 1210–1226. MR3621018 https://doi.org/10.1002/sim.7223

Key words and phrases. Bayesian inference, commensurate prior, network meta-analysis, randomized controlled trial, single-arm trial.

EGGER, M., DAVEY SMITH, G. and ALTMAN, D. G., eds. (2001) *Systematic Reviews in Health Care*. BMJ Publishing Group.

GAMALO, M. A., TIWARI, R. C. and LAVANGE, L. M. (2013). Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products. *Pharmaceutical Statistics* **13** 25–40.

GAMALO-SIEBERS, M., SAVIC, J., BASU, C., ZHAO, X., GOPALAKRISHNAN, M., GAO, A., SONG, G., BAYGANI, S., THOMPSON, L. et al. (2017). Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation. *Pharm. Stat.* **16** 232–249.

GEISSER, S. and EDDY, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74** 153–160. MR0529531

HANSON, T. E., BRANSCUM, A. J. and JOHNSON, W. O. (2011). Predictive comparison of joint longitudinal-survival modeling: A case study illustrating competing approaches. *Lifetime Data Anal.* **17** 3–28. MR2764577 https://doi.org/10.1007/s10985-010-9162-0

HOBBS, B. P., SARGENT, D. J. and CARLIN, B. P. (2012). Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Anal.* **7** 639–673. MR2981631 https://doi.org/10.1214/12-BA722

HOBBS, B. P., CARLIN, B. P., MANDREKAR, S. J. and SARGENT, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* **67** 1047–1056. MR2829239 https://doi.org/10.1111/j.1541-0420.2011.01564.x

HONG, H., FU, H. and CARLIN, B. P. (2018). Power and commensurate priors for synthesizing aggregate and individual patient level data in network meta-analysis. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 1047–1069. MR3832263 https://doi.org/10.1111/rssc.12275

HONG, H., CHU, H., ZHANG, J. and CARLIN, B. P. (2016a). A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Res. Synth. Methods* **7** 6–22.

HONG, H., CHU, H., ZHANG, J. and CARLIN, B. P. (2016b). Rejoinder to the discussion of "a Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons," by S. Dias and A.E. Ades. *Res. Synth. Methods* **7** 29–33.

HUEBER, W., SANDS, B. E., LEWITZKY, S., VANDEMEULEBROECKE, M., REINISCH, W., HIGGINS, P. D. R., WEHKAMP, J., FEAGAN, B. G., YAO, M. D. et al. (2012). Secukinumab, a human anti-IL-17A monoclonal antibody, for moderate to severe Crohn's disease: Unexpected results of a randomised, double-blind placebo-controlled trial. *Gut* **61** 1693–1700.

IBRAHIM, J. G. and CHEN, M.-H. (2000). Power prior distributions for regression models. *Statist. Sci.* **15** 46–60. MR1842236 https://doi.org/10.1214/ss/1009212673

IBRAHIM, J. G., CHEN, M.-H., GWON, Y. and CHEN, F. (2015). The power prior: Theory and applications. *Stat. Med.* **34** 3724–3749. MR3422144 https://doi.org/10.1002/sim.6728

JACKSON, D., BARRETT, J. K., RICE, S., WHITE, I. R. and HIGGINS, J. P. T. (2014). A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. *Stat. Med.* **33** 3639–3654. MR3260651 https://doi.org/10.1002/sim.6188

JAFF, M. R., NELSON, T., FERKO, N., MARTINSON, M., ANDERSON, L. H. and HOLLMANN, S. (2017). Endovascular interventions for femoropopliteal peripheral artery disease: A network meta-analysis of current technologies. *J. Vasc. Interv. Radiol.* **28** 1617–1627.

JOHNSON, D. B., CHANDRA, S. and SOSMAN, J. A. (2018). Immune checkpoint inhibitor toxicity in 2018. *JAMA* **320** 1702–1703. https://doi.org/10.1001/jama.2018.13995

KAIZER, A. M., HOBBS, B. P. and KOOPMEINERS, J. S. (2018). A multi-source adaptive platform design for testing sequential combinatorial therapeutic strategies. *Biometrics* **74** 1082–1094. MR3860729 https://doi.org/10.1111/biom.12841

KAIZER, A. M., KOOPMEINERS, J. S. and HOBBS, B. P. (2018). Bayesian hierarchical modeling based on multisource exchangeability. *Biostatistics* **19** 169–184. MR3799610 https://doi.org/10.1093/biostatistics/kxx031

KONTOPANTELIS, E., SPRINGATE, D. A. and REEVES, D. (2013). A re-analysis of the Cochrane Library data: The dangers of unobserved heterogeneity in meta-analyses. *PLoS ONE* **8** e69930. https://doi.org/10.1371/journal.pone.0069930

LEAHY, J., THOM, H., JANSEN, J. P., GRAY, E., O'LEARY, A., WHITE, A. and WALSH, C. (2019). Incorporating single-arm evidence into a network meta-analysis using aggregate level matching: Assessing the impact. *Stat. Med.* **38** 2505–2523. MR3962125 https://doi.org/10.1002/sim.8139

LI, Z. and BEGG, C. B. (1994). Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. *J. Amer. Statist. Assoc.* **89** 1523–1527. MR1310241

LIN, L., CHU, H. and HODGES, J. S. (2016). Sensitivity to excluding treatments in network meta-analysis. *Epidemiology* **27** 562–569.

LIN, L., ZHANG, J., HODGES, J. S. and CHU, H. (2017). Performing arm-based network meta-analysis in R with the pcnetmeta package. *J. Stat. Softw.* **80** 1–25.

LU, G. and ADES, A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Stat. Med.* **23** 3105–3124.
LU, G. and ADES, A. E. (2006b). Assessing evidence inconsistency in mixed treatment comparisons. *J. Amer. Statist. Assoc.* **101** 447–459. MR2256166 https://doi.org/10.1198/016214505000001302
LU, G. and ADES, A. E. (2009). Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics* **10** 792–805.
LUNN, D., JACKSON, C., BEST, N., SPIEGELHALTER, D. and THOMAS, A. (2010). *The BUGS Book*. Taylor & Francis, London.
MATHES, T. and KUSS, O. (2018). A comparison of methods for meta-analysis of a small number of studies with binary outcomes. *Res. Synth. Methods* **9** 366–381.
MORRIS, T. P., WHITE, I. R. and CROWTHER, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* **38** 2074–2102. MR3937487 https://doi.org/10.1002/sim.8086
MURRAY, T. A., HOBBS, B. P. and CARLIN, B. P. (2015). Combining nonexchangeable functional or survival data sources in oncology using generalized mixture commensurate priors. *Ann. Appl. Stat.* **9** 1549–1570. MR3418735 https://doi.org/10.1214/15-AOAS840
NIKOLAKOPOULOU, A., CHAIMANI, A., VERONIKI, A. A., VASILIADIS, H. S., SCHMID, C. H. and SALANTI, G. (2014). Characteristics of networks of interventions: A description of a database of 186 published networks. *PLoS ONE* **9** e86754. https://doi.org/10.1371/journal.pone.0086754
PHILLIPPO, D. M., DIAS, S., ADES, A. E. and WELTON, N. J. (2020). Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study. *Stat. Med.* **39** 4885–4911. MR4190952 https://doi.org/10.1002/sim.8759
RÖVER, C., WANDEL, S. and FRIEDE, T. (2019). Model averaging for robust extrapolation in evidence synthesis. *Stat. Med.* **38** 674–694. MR3902606 https://doi.org/10.1002/sim.7991
SALANTI, G., ADES, A. E. and IOANNIDIS, J. P. A. (2011). Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: An overview and tutorial. *J. Clin. Epidemiol.* **64** 163–171.
SCHMIDLI, H., GSTEIGER, S., ROYCHOUDHURY, S., O'HAGAN, A., SPIEGELHALTER, D. and NEUENSCHWANDER, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* **70** 1023–1032. MR3295763 https://doi.org/10.1111/biom.12242
SCHMITZ, S., MAGUIRE, Á., MORRIS, J., RUGGERI, K., HALLER, E., KUHN, I., LEAHY, J., HOMER, N., KHAN, A. et al. (2018). The use of single armed observational data to closing the gap in otherwise disconnected evidence networks: A network meta-analysis in multiple myeloma. *BMC Med. Res. Methodol.* **18** 66.
SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. MR1979380 https://doi.org/10.1111/1467-9868.00353
THOM, H. H. Z., CAPKUN, G., CERULLI, A., NIXON, R. M. and HOWARD, L. S. (2015). Network meta-analysis combining individual patient and aggregate data from a mixture of study designs with an application to pulmonary arterial hypertension. *BMC Med. Res. Methodol.* **15** 34. https://doi.org/10.1186/s12874-015-0007-0
TURNER, R. M., DOMÍNGUEZ-ISLAS, C. P., JACKSON, D., RHODES, K. M. and WHITE, I. R. (2019). Incorporating external evidence on between-trial heterogeneity in network meta-analysis. *Stat. Med.* **38** 1321–1335. MR3920617 https://doi.org/10.1002/sim.8044
WANG, Z., LIN, L., HODGES, J. S. and CHU, H. (2020a). The impact of covariance priors on arm-based Bayesian network meta-analyses with binary outcomes. *Stat. Med.* **39** 2883–2900. MR4151887 https://doi.org/10.1002/sim.8580
WANG, Z., LIN, L., ZHAO, S. and CHU, H. (2020b). Nmaplateplot: The Plate Plot for Network Meta-Analysis Results. R package version 1.0.0.
WANG, Z., LIN, L., HODGES, J. S., MACLEHOSE, R. and CHU, H. (2021a). A variance shrinkage method improves arm-based Bayesian network meta-analysis. *Stat. Methods Med. Res.* **30** 151–165. MR4216852 https://doi.org/10.1177/0962280220945731
WANG, Z., LIN, L., MURRAY, T., HODGES, J. S. and CHU, H. (2021b). Supplement to "Bridging randomized controlled trials and single-arm trials using commensurate priors in arm-based network meta-analysis." https://doi.org/10.1214/21-AOAS1469SUPPA, https://doi.org/10.1214/21-AOAS1469SUPPB
WELTON, N. J., SUTTON, A. J., COOPER, N. J. and ABRAMS, K. R. (2012). *Evidence Synthesis for Decision Making in Healthcare*. Wiley, New York.
WHITE, I. R., BARRETT, J. K., JACKSON, D. and HIGGINS, J. P. T. (2012). Consistency and inconsistency in network meta-analysis: Model estimation using multivariate meta-regression. *Res. Synth. Methods* **3** 111–125.
WHITE, I. R., TURNER, R. M., KARAHALIOS, A. and SALANTI, G. (2019). A comparison of arm-based and contrast-based models for network meta-analysis. *Stat. Med.* **38** 5197–5213. MR4032399 https://doi.org/10.1002/sim.8360

XU, C., CHEN, Y.-P., DU, X.-J., LIU, J.-Q., HUANG, C.-L., CHEN, L., ZHOU, G.-Q., LI, W.-F., MAO, Y.-P. et al. (2018). Comparative safety of immune checkpoint inhibitors in cancer: Systematic review and network meta-analysis. *BMJ* k4226.

ZEGER, S. L., LIANG, K.-Y. and ALBERT, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44** 1049–1060. MR0980999 https://doi.org/10.2307/2531734

ZHANG, J., CARLIN, B. P., NEATON, J. D., SOON, G. G., NIE, L., KANE, R., VIRNIG, B. A. and CHU, H. (2014). Network meta-analysis of randomized clinical trials: Reporting the proper summaries. *Clin. Trials* **11** 246–262.

ZHANG, J., CHU, H., HONG, H., VIRNIG, B. A. and CARLIN, B. P. (2017a). Bayesian hierarchical models for network meta-analysis incorporating nonignorable missingness. *Stat. Methods Med. Res.* **26** 2227–2243. MR3712230 https://doi.org/10.1177/0962280215596185

ZHANG, D., CHEN, M.-H., IBRAHIM, J. G., BOYE, M. E. and SHEN, W. (2017b). Bayesian model assessment in joint modeling of longitudinal and survival data with applications to cancer clinical trials. *J. Comput. Graph. Statist.* **26** 121–133. MR3610413 https://doi.org/10.1080/10618600.2015.1117472

ZHANG, J., KO, C.-W., NIE, L., CHEN, Y. and TIWARI, R. (2019). Bayesian hierarchical methods for meta-analysis combining randomized-controlled and single-arm studies. *Stat. Methods Med. Res.* **28** 1293–1310. MR3941077 https://doi.org/10.1177/0962280218754928

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1788–1807
https://doi.org/10.1214/21-AOAS1449

INFORMATION CONTENT OF HIGH-ORDER ASSOCIATIONS OF THE HUMAN GUT MICROBIOTA NETWORK

BY WESTON D. VILES[1,*], JULIETTE C. MADAN[1,†], HONGZHE LI[2], MARGARET R. KARAGAS[1,‡] AND ANNE G. HOEN[1,§]

[1]*Geisel School of Medicine, Dartmouth College,* [*]*weston.viles@maine.edu;* [†]*juliette.c.madan@hitchcock.org;* [‡]*margaret.r.karagas@dartmouth.edu;* [§]*anne.g.hoen@dartmouth.edu*

[2]*Perelman School of Medicine, University of Pennsylvania, hongzhe@upenn.edu*

The human gastrointestinal tract is an environment that hosts an ecosystem of microorganisms essential to human health. Vital biological processes emerge from fundamental inter- and intraspecies molecular interactions that influence the assembly and composition of the gut microbiota ecology. Here, we quantify the complexity of the ecological relationships within the human infant gut microbiota ecosystem as a function of the information contained in the nonlinear associations of a sequence of increasingly specified maximum entropy representations of the system. Our paradigm frames the ecological state, in terms of the presence or absence of individual microbial ecological units that are identified by amplicon sequence variants (ASV) in the gut microenvironment, as a function of both the ecological states of its neighboring units and, in a departure from standard graphical model representations, the associations among the units within its neighborhood. We characterize the order of the system based on the relative quantity of statistical information encoded by high-order statistical associations of the infant gut microbiota.

REFERENCES

AAS, J. A., PASTER, B. J., STOKES, L. N., OLSEN, I. and DEWHIRST, F. E. (2005). Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* **43** 5721–5732.

ABRAMS, P. A. (1983). Arguments in favor of higher order interactions. *Amer. Nat.* **121** 887–891.

AITCHISON, J. (1981). A new approach to null correlations of proportions. *J. Int. Assoc. Math. Geol.* **13** 175–189. MR0613760 https://doi.org/10.1007/BF01031393

ALON, U. (2007). Network motifs: Theory and experimental approaches. *Nat. Rev. Genet.* **8** 450–461.

ANTOSCA, K., HOEN, A. G., PALYS, T., HILLIARD, M., MORRISON, H. G., COKER, M., MADAN, J. and KARAGAS, M. R. (2020). Reliability of stool microbiome methods for DNA yields and sequencing among infants and young children. *Microbiologyopen* **9** e1018. https://doi.org/10.1002/mbo3.1018

BÄCKHED, F., LEY, R. E., SONNENBURG, J. L., PETERSON, D. A. and GORDON, J. I. (2005). Host-bacterial mutualism in the human intestine. *Science* **307** 1915–1920.

BAIREY, E., KELSIC, E. D. and KISHONY, R. (2016). High-order species interactions shape ecosystem diversity. *Nat. Commun.* **7** 12285. https://doi.org/10.1038/ncomms12285

BAR-MASSADA, A. (2015). Complex relationships between species niches and environmental heterogeneity affect species co-occurrence patterns in modelled and real communities. *Proc. R. Soc. Lond., B Biol. Sci.* **282** 20150927. https://doi.org/10.1098/rspb.2015.0927

BARBOUR, A. D., HOLST, L. and JANSON, S. (1992). *Poisson Approximation. Oxford Studies in Probability* **2**. The Clarendon Press, Oxford University Press, New York. Oxford Science Publications. MR1163825

BÄUMLER, A. J. and SPERANDIO, V. (2016). Interactions between the microbiota and pathogenic bacteria in the gut. *Nature* **535** 85–93. https://doi.org/10.1038/nature18849

BECKERMAN, A. P., URIARTE, M. and SCHMITZ, O. J. (1997). Experimental evidence for a behavior-mediated trophic cascade in a terrestrial food chain. *Proc. Natl. Acad. Sci. USA* **94** 10735–10738.

BILLICK, I. and CASE, T. J. (1994). Higher order interactions in ecological communities: What are they and how can they be detected? *Ecology* **75** 1530–1543.

BOTEV, Z. I. and KROESE, D. P. (2011). The generalized cross entropy method, with applications to probability density estimation. *Methodol. Comput. Appl. Probab.* **13** 1–27. MR2755130 https://doi.org/10.1007/s11009-009-9133-7

Key words and phrases. Gut microbiota, ecosystem, high-order associations, information, relative entropy.

CALLAHAN, B. J., MCMURDIE, P. J. and HOLMES, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11** 2639–2643. https://doi.org/10.1038/ismej.2017.119

CALLAHAN, B. J., MCMURDIE, P. J., ROSEN, M. J., HAN, A. W., JOHNSON, A. J. A. and HOLMES, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13** 581–583.

CASE, T. J. and BENDER, E. A. (1981). Testing for higher order interactions. *Amer. Nat.* **118** 920–929. MR0647503 https://doi.org/10.1086/283885

CHEN, I., KELKAR, Y. D., GU, Y., ZHOU, J., QIU, X. and WU, H. (2017). High-dimensional linear state space models for dynamic microbial interaction networks. *PLoS ONE* **12** e0187822.

DURBIN, J. and KOOPMAN, S. J. (2001). *Time Series Analysis by State Space Methods*. *Oxford Statistical Science Series* **24**. Oxford Univ. Press, Oxford. MR1856951

EDGAR, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26** 2460–2461.

EDGAR, R. C., HAAS, B. J., CLEMENTE, J. C., QUINCE, C. and KNIGHT, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27** 2194–2200.

FAUST, K. and RAES, J. (2012). Microbial interactions: From networks to models. *Nat. Rev., Microbiol.* **10** 538–550.

FAUST, K., SATHIRAPONGSASUTI, J. F., IZARD, J., SEGATA, N., GEVERS, D., RAES, J. and HUTTENHOWER, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8** e1002606.

FISHER, J. C., EREN, A. M., GREEN, H. C., SHANKS, O. C., MORRISON, H. G., VINEIS, J. H., SOGIN, M. L., MCLELLAN, S. L. and SCHAFFNER, D. W. (2015). Comparison of sewage and animal fecal microbiomes by using oligotyping reveals potential human fecal indicators in multiple taxonomic groups. *Appl. Environ. Microbiol.* **81** 7023–7033.

FREILICH, S., KREIMER, A., MEILIJSON, I., GOPHNA, U., SHARAN, R. and RUPPIN, E. (2010). The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Res.* **38** 3857–3868.

FRIEDMAN, J., HIGGINS, L. M. and GORE, J. (2017). Community structure follows simple assembly rules in microbial microcosms. *Nat. Ecol. Evol.* **1** 0109.

GOULD, A. L., ZHANG, V., LAMBERTI, L., JONES, E. W., OBADIA, B., KORASIDIS, N., GAVRYUSHKIN, A., CARLSON, J. M., BEERENWINKEL, N. et al. (2018). Microbiome interactions shape host fitness. *Proc. Natl. Acad. Sci. USA* **115** E11951–E11960. https://doi.org/10.1073/pnas.1809349115

GRILLI, J., BARABÁS, G., MICHALSKA-SMITH, M. J. and ALLESINA, S. (2017). Higher-order interactions stabilize dynamics in competitive network models. *Nature* **548** 210–213. https://doi.org/10.1038/nature23273

HAQUE, S. Z. and HAQUE, M. (2017). The ecological community of commensal, symbiotic, and pathogenic gastrointestinal microorganisms—an appraisal. *Clinical and Experimental Gastroenterology* **10** 91–103.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

HUSE, S. M., YOUNG, V. B., MORRISON, H. G., ANTONOPOULOS, D. A., KWON, J., DALAL, S., ARRIETA, R., HUBERT, N. A., SHEN, L. et al. (2014). Comparison of brush and biopsy sampling methods of the ileal pouch for assessment of mucosa-associated microbiota of human subjects. *Microbiome* **2** 5.

IVES, A. R. and CARPENTER, S. R. (2007). Stability and diversity of ecosystems. *Science* **317** 58–62. https://doi.org/10.1126/science.1133258

JAYNES, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.* (2) **106** 620–630. MR0087305

KÅHRSTRÖM, C. T., PARIENTE, N. and WEISS, U. (2016). Intestinal microbiota in health and disease. *Nature* **535** 47.

KAY, G. M., TULLOCH, A., BARTON, P. S., CUNNINGHAM, S. A., DRISCOLL, D. A. and LINDENMAYER, D. B. (2018). Species co-occurrence networks show reptile community reorganization under agricultural transformation. *Ecography* **41** 13–125.

KELSIC, E. D., ZHAO, J., VETSIGIAN, K. and KISHONY, R. (2015). Counteraction of antibiotic production and degradation stabilizes microbial communities. *Nature* **521** 516–519.

KIEFER, J. (1953). Sequential minimax search for a maximum. *Proc. Amer. Math. Soc.* **4** 502–506. MR0055639 https://doi.org/10.2307/2032161

KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* **22** 79–86. MR0039968 https://doi.org/10.1214/aoms/1177729694

LAYEGHIFARD, M., HWANG, D. M. and GUTTMAN, D. S. (2017). Disentangling interactions in the microbiome: A network perspective. *Trends Microbiol.* **25** 217–228.

LEVINE, J., BASCOMPTE, J., ADLER, P. and ALLESINA, S. (2017). Beyond pairwise coexistence: Biodiversity maintenance in complex ecological communities. *Nature* **546** 56–64.

MACKENZIE, D. L., BAILEY, L. L. and NICHOLS, J. D. (2004). Investigating species co-occurrence patterns when species are detected imperfectly. *J. Anim. Ecol.* **73** 546–555.

MANDAKOVIC, D., ROJAS, C., MALDONADO, J., LATORRE, M., TRAVISANY, D., DELAGE, E., BIHOUÉE, A., JEAN, G., DÍAZ, F. P. et al. (2018). Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Sci. Rep.* **8** 5875.

MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D. and ALON, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* **298** 824–827.

MORUETA-HOLME, N., BLONDER, B., SANDEL, B. S., MCGILL, B. J., PEET, R. K., OTT, J. E., VIOLLE, C., ENQUIST, B. J., JORGENSEN, P. M. et al. (2016). A network approach for inferring species associations from co-occurrence data. *Ecography* **39** 1139–1150.

POISOT, T., STOUFFER, D. B. and GRAVEL, D. (2015). Beyond species: Why ecological interactions vary through space and time. *Oikos* **124** 243–251.

POUDEL, R., JUMPPONEN, A., SCHLATTER, D. C., PAULITZ, T. C., GARDENER, B. B. M., KINKEL, L. L. and GARRETT, K. A. (2016). Microbiome networks: A systems framework for identifying candidate microbial assemblages for disease management. *Phytopathology* **106** 1083–1096.

RAMETTE, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol.* **62** 142–160.

ROTHSTEIN, J. (1952). Organization and entropy. *J. Appl. Phys.* **23** 1281–1282.

SHOAIE, S., KARLSSON, F., MARDINOGLU, A., NOOKAEW, I., BORDEL, S. and NIELSEN, J. (2013). Understanding the interactions between bacteria in the human gut through metabolic modeling. *Sci. Rep.* **3** 2532. https://doi.org/10.1038/srep02532

SINGH, S. B., MADAN, J., COKER, M., HOEN, A., BAKER, E. R., KARAGAS, M. R. and MUELLER, N. T. (2020). Does birth mode modify associations of maternal pre-pregnancy BMI and gestational weight gain with the infant gut microbiome? *Int. J. Obes.* **44** 23–32. https://doi.org/10.1038/s41366-018-0273-0

TROSVIK, P., RUDI, K., STRAETKVERN, K. O., JAKOBSEN, K. S., NAES, T. and STENSETH, N. C. (2010). Web of ecological interactions in an experimental gut microbiota. *Environ. Microbiol.* **10** 2677–2687.

TSAI, K.-N., LIN, S.-H., LIU, W.-C. and WANG, D. (2015). Inferring microbial interaction network from microbiome data using RMN algorithm. *BMC Syst. Biol.* **9** 54.

VAN DEN BERGH, M. R., BIESBROEK, G., ROSSEN, J. W. A., DE STEENHUIJSEN PITERS, W. A. A., BOSCH, A. A. T. M., VAN GILS, E. J. M., WANG, X., BOONACKER, C. W. B., VEENHOVEN, R. H. et al. (2012). Associations between pathogens in the upper respiratory tract of young children: Interplay between viruses and bacteria. *PLoS ONE* **7** e4771.

VILES, W. D., MADAN, J. C., LI, H., KARAGAS, M. R. and HOEN, A. G. (2021). Supplement to "Information content of high-order associations of the human gut microbiota network." https://doi.org/10.1214/21-AOAS1449SUPP

WOOTTON, J. T. (1994). The nature and consequences of indirect effects in ecological communities. *Ann. Rev. Ecolog. Syst.* **25** 443–466.

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1808–1830
https://doi.org/10.1214/21-AOAS1458

RADIOHEAD: RADIOGENOMIC ANALYSIS INCORPORATING TUMOR HETEROGENEITY IN IMAGING THROUGH DENSITIES

BY SHARIQ MOHAMMED[1,*], KARTHIK BHARATH[2], SEBASTIAN KURTEK[3], ARVIND RAO[1,†] AND VEERABHADRAN BALADANDAYUTHAPANI[1,‡]

[1]*Department of Biostatistics, Department of Computational Medicine and Bioinformatics, University of Michigan,* [*]*shariqm@umich.edu;* [†]*ukarvind@umich.edu;* [‡]*veerab@umich.edu*

[2]*School of Mathematical Sciences, University of Nottingham, karthik.bharath@nottingham.ac.uk*

[3]*Department of Statistics, The Ohio State University, kurtek.1@stat.osu.edu*

Recent technological advancements have enabled detailed investigation of associations between the molecular architecture and tumor heterogeneity through multisource integration of radiological imaging and genomic (radiogenomic) data. In this paper we integrate and harness radiogenomic data in patients with lower grade gliomas (LGG), a type of brain cancer, in order to develop a regression framework called RADIOHEAD (RADIOgenomic analysis incorporating tumor HEterogeneity in imAging through Densities) to identify radiogenomic associations. Imaging data is represented through voxel-intensity probability density functions of tumor subregions obtained from multimodal magnetic resonance imaging and genomic data through molecular signatures in the form of pathway enrichment scores corresponding to their gene expression profiles. Employing a Riemannian-geometric framework for principal component analysis on the set of probability density functions, we map each probability density to a vector of principal component scores which are then included as predictors in a Bayesian regression model with the pathway enrichment scores as the response. Variable selection compatible with the grouping structure amongst the predictors induced through the tumor subregions is carried out under a group spike-and-slab prior. A Bayesian false discovery rate mechanism is then used to infer significant associations based on the posterior distribution of the regression coefficients. Our analyses reveal several pathways relevant to LGG etiology (such as synaptic transmission, nerve impulse and neurotransmitter pathways) to have significant associations with the corresponding imaging-based predictors.

REFERENCES

ANDERSEN, M. R., WINTHER, O. and HANSEN, L. K. (2014). Bayesian inference for structured spike and slab priors. In *Advances in Neural Information Processing Systems* 1745–1753.

BAEK, H. J., KIM, H. S., KIM, N., CHOI, Y. J. and KIM, Y. J. (2012). Percent change of perfusion skewness and kurtosis: A potential imaging biomarker for early treatment response in patients with newly diagnosed glioblastomas. *Radiology* **264** 834–843.

BAKAS, S., ZENG, K., SOTIRAS, A., RATHORE, S., AKBARI, H., GAONKAR, B., ROZYCKI, M., PATI, S. and DAVATZIKOS, C. (2015). GLISTRboost: Combining multimodal MRI segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. In *International Workshop on Brainlesion*: *Glioma*, *Multiple Sclerosis*, *Stroke and Traumatic Brain Injuries* 144–155. Springer, Berlin.

BAKAS, S., AKBARI, H., SOTIRAS, A., BILELLO, M., ROZYCKI, M., KIRBY, J. S., FREYMANN, J. B., FARAHANI, K. and DAVATZIKOS, C. (2017a). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* **4** 170117.

BAKAS, S., AKBARI, H., SOTIRAS, A., BILELLO, M., ROZYCKI, M., KIRBY, J., FREYMANN, J., FARAHANI, K. and DAVATZIKOS, C. (2017b). Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection. Available at https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF.

Key words and phrases. Fisher–Rao metric, group spike-and-slab, principal component analysis, radiogenomic associations.

BHATTACHARYYA, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **35** 99–109. MR0010358

BILIOTTI, L. and MERCURI, F. (2017). Riemannian Hilbert manifolds. In *Hermitian–Grassmannian Submanifolds. Springer Proc. Math. Stat.* **203** 261–271. Springer, Singapore. MR3710849 https://doi.org/10.1007/978-981-10-5556-0_2

BROCKE, K. S., STAUFNER, C., LUKSCH, H., GEIGER, K. D., STEPULAK, A., MARZAHN, J., SCHACKERT, G., TEMME, A. and IKONOMIDOU, C. (2010). Glutamate receptors in pediatric tumors of the central nervous system. *Cancer Biol. Ther.* **9** 455–468.

CECCARELLI, M., BARTHEL, F. P., MALTA, T. M., SABEDOT, T. S., SALAMA, S. R., MURRAY, B. A., MOROZOVA, O., NEWTON, Y., RADENBAUGH, A. et al. (2016). Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164** 550–563.

ČENCOV, N. N. (1982). *Statistical Decision Rules and Optimal Inference. Translations of Mathematical Monographs* **53**. Amer. Math. Soc., Providence, RI. MR0645898

CHERRY, A. E. and STELLA, N. (2014). G protein-coupled receptors as oncogenic signals in glioma: Emerging therapeutic avenues. *Neuroscience* **278** 222–236.

CLARK, K., VENDT, B., SMITH, K., FREYMANN, J., KIRBY, J., KOPPEL, P., MOORE, S., PHILLIPS, S., MAFFITT, D. et al. (2013). The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *J. Digit. Imaging* **26** 1045–1057.

DRYDEN, I. L. and MARDIA, K. V. (1998). *Statistical Shape Analysis. Wiley Series in Probability and Statistics: Probability and Statistics*. Wiley, Chichester. MR1646114

FISHBEIN, L., LESHCHINER, I., WALTER, V., DANILOVA, L., ROBERTSON, A. G., JOHNSON, A. R., LICHTENBERG, T. M., MURRAY, B. A., GHAYEE, H. K. et al. (2017). Comprehensive molecular characterization of pheochromocytoma and paraganglioma. *Cancer Cell* **31** 181–193. https://doi.org/10.1016/j.ccell.2017.01.001

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GENTLEMAN, R. C., CAREY, V. J., BATES, D. M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y. et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **5** R80.

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* 339–373.

HÄNZELMANN, S., CASTELO, R. and GUINNEY, J. (2013). GSVA: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* **14** 7. https://doi.org/10.1186/1471-2105-14-7

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science* 382–401.

ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158 https://doi.org/10.1214/009053604000001147

JUST, N. (2011). Histogram analysis of the microvasculature of intracerebral human and murine glioma xenografts. *Magn. Reson. Med.* **65** 778–789.

JUST, N. (2014). Improving tumour heterogeneity MRI assessment with histograms. *Br. J. Cancer* **111** 2205–2213. https://doi.org/10.1038/bjc.2014.512

KARCHER, H. (1977). Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.* **30** 509–541. MR0442975 https://doi.org/10.1002/cpa.3160300502

KASS, R. E. and VOS, P. W. (2011). *Geometrical Foundations of Asymptotic Inference* **908**. Wiley, New York.

KURTEK, S. and BHARATH, K. (2015). Bayesian sensitivity analysis with the Fisher–Rao metric. *Biometrika* **102** 601–616. MR3394278 https://doi.org/10.1093/biomet/asv026

LIBERZON, A., SUBRAMANIAN, A., PINCHBACK, R., THORVALDSDÓTTIR, H., TAMAYO, P. and MESIROV, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27** 1739–1740.

LIU, Y., ZHOU, Y. and ZHU, K. (2012). Inhibition of glioma cell lysosome exocytosis inhibits glioma invasion. *PLoS ONE* **7** e45910.

MARUSYK, A., ALMENDRO, V. and POLYAK, K. (2012). Intra-tumour heterogeneity: A looking glass for cancer? *Nat. Rev. Cancer* **12** 323–334. https://doi.org/10.1038/nrc3261

MOHAMMED, S., BHARATH, K., KURTEK, S., RAO, A. and BALADANDAYUTHAPANI, V. (2021a). Supplement to "RADIOHEAD: Radiogenomic analysis incorporating tumor heterogeneity in imagine through densities." https://doi.org/10.1214/21-AOAS1458SUPPA

MOHAMMED, S., BHARATH, K., KURTEK, S., RAO, A. and BALADANDAYUTHAPANI, V. (2021b). Code for "RADIOHEAD: Radiogenomic analysis incorporating tumor heterogeneity in imagine through densities." https://doi.org/10.1214/21-AOAS1458SUPPB

MORRIS, J. S. and CARROLL, R. J. (2006). Wavelet-based functional mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 179–199. MR2188981 https://doi.org/10.1111/j.1467-9868.2006.00539.x

MORRIS, J. S., BROWN, P. J., HERRICK, R. C., BAGGERLY, K. A. and COOMBES, K. R. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics* **64** 479–489.

MÜLLER, P., PARMIGIANI, G., ROBERT, C. and ROUSSEAU, J. (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *J. Amer. Statist. Assoc.* **99** 990–1001. MR2109489 https://doi.org/10.1198/016214504000001646

NOUSHMEHR, H., WEISENBERGER, D. J., DIEFES, K., PHILLIPS, H. S., PUJARA, K., BERMAN, B. P., PAN, F., PELLOSKI, C. E., SULMAN, E. P. et al. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17** 510–522.

OMBAO, H., LINDQUIST, M., THOMPSON, W. and ASTON, J. (2016). *Handbook of Neuroimaging Data Analysis*. CRC Press, Boca Raton.

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2168993

RAO, C. R. (1992). Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in Statistics* 235–247. Springer, Berlin.

SAHA, A., BANERJEE, S., KURTEK, S., NARANG, S., LEE, J., RAO, G., MARTINEZ, J., BHARATH, K., RAO, A. U. et al. (2016). DEMARCATE: Density-based magnetic resonance image clustering for assessing tumor heterogeneity in cancer. *NeuroImage Clin.* **12** 132–143.

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

SHINOHARA, R. T., SWEENEY, E. M., GOLDSMITH, J., SHIEE, N., MATEEN, F. J., CALABRESI, P. A., JARSO, S., PHAM, D. L., REICH, D. S. et al. (2014). Statistical normalization techniques for magnetic resonance imaging. *NeuroImage Clin.* **6** 9–19.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR0848134 https://doi.org/10.1007/978-1-4899-3324-9

SONG, Y. S., CHOI, S. H., PARK, C.-K., YI, K. S., LEE, W. J., YUN, T. J., KIM, T. M., LEE, S.-H., KIM, J.-H. et al. (2013). True progression versus pseudoprogression in the treatment of glioblastomas: A comparison study of normalized cerebral blood volume and apparent diffusion coefficient by histogram analysis. *Korean J. Radiol.* **14** 662–672.

SRIVASTAVA, A., JERMYN, I. H. and JOSHI, S. H. (2007). Riemannian analysis of probability density functions with applications in vision. In *IEEE Conference on Computer Vision and Pattern Recognition* 1–8.

STOREY, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035. MR2036398 https://doi.org/10.1214/aos/1074290335

VASAIKAR, S. V., STRAUB, P., WANG, J. and ZHANG, B. (2017). LinkedOmics: Analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* **46** D956–D963.

VENKATESH, H. S., JOHUNG, T. B., CARETTI, V., NOLL, A., TANG, Y., NAGARAJA, S., GIBSON, E. M., MOUNT, C. W., POLEPALLI, J. et al. (2015). Neuronal activity promotes glioma growth through neuroligin-3 secretion. *Cell* **161** 803–816.

VENNETI, S. and HUSE, J. T. (2015). The evolving molecular genetics of low-grade glioma. *Adv. Anat. Pathol.* **22** 94–101. https://doi.org/10.1097/PAP.0000000000000049

VERHAAK, R. G., HOADLEY, K. A., PURDOM, E., WANG, V., QI, Y., WILKERSON, M. D., MILLER, C. R., DING, L., GOLUB, T. et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17** 98–110.

VERHAAK, R. G. W., COOPER, L. A. D., SALAMA, S. S., ALDAPE, K., YUNG, W. K. A. and BRAT, D. J. (2014). Abstract 936: Comprehensive and integrative genomic characterization of diffuse lower grade gliomas. *Cancer Res.* **74** 936–936.

WANG, R., GURGUIS, C. I., GU, W., KO, E. A., LIM, I., BANG, H., ZHOU, T. and KO, J.-H. (2015). Ion channel gene expression predicts survival in glioma patients. *Sci. Rep.* **5** 11593.

XU, X. and GHOSH, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Anal.* **10** 909–936. MR3432244 https://doi.org/10.1214/14-BA929

YANG, X. and NARISETTY, N. N. (2020). Consistent group selection with Bayesian high dimensional modeling. *Bayesian Anal.* **15** 909–935. MR4132654 https://doi.org/10.1214/19-BA1178

YANG, H., BALADANDAYUTHAPANI, V., RAO, A. U. K. and MORRIS, J. S. (2020). Quantile function on scalar regression analysis for distributional data. *J. Amer. Statist. Assoc.* **115** 90–106. MR4078447 https://doi.org/10.1080/01621459.2019.1609969

ZHANG, L., BALADANDAYUTHAPANI, V., MALLICK, B. K., MANYAM, G. C., THOMPSON, P. A., BONDY, M. L. and DO, K.-A. (2014). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 595–620. MR3258055 https://doi.org/10.1111/rssc.12053

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1831–1849
https://doi.org/10.1214/21-AOAS1460

ASSESSING THE RELIABILITY OF WIND POWER OPERATIONS UNDER A CHANGING CLIMATE WITH A NON-GAUSSIAN BIAS CORRECTION

BY JIACHEN ZHANG[1,*], PAOLA CRIPPA[2], MARC G. GENTON[3] AND STEFANO CASTRUCCIO[1,†]

[1]*Department of Applied and Computational Mathematics and Statistics, University of Notre Dame,* [*]*jzhang19@nd.edu;* [†]*scastruc@nd.edu*

[2]*Department of Civil and Environmental Engineering and Geosciences, University of Notre Dame, pcrippa@nd.edu*

[3]*Statistics Program, King Abdullah University of Science and Technology, marc.genton@kaust.edu.sa*

Facing increasing societal and economic pressure, many countries have established strategies to develop renewable energy portfolios whose penetration in the market can alleviate the dependence on fossil fuels. In the case of wind, there is a fundamental question related to the resilience and hence profitability of future wind farms to a changing climate, given that current wind turbines have lifespans of up to 30 years. In this work we develop a new non-Gaussian method to adjust assimilated observational data to simulations and to estimate future wind, predicated on a trans-Gaussian transformation and a clusterwise minimization of the Kullback–Leibler divergence. Future winds abundance will be determined for Saudi Arabia, a country with a recently established plan to develop a portfolio of up to 16 GW of wind energy. Further, we estimate the change in profits over future decades using additional high-resolution simulations, an improved method for vertical wind extrapolation and power curves from a collection of popular wind turbines. We find an overall increase in daily profit of $272,000 for the wind energy market for the optimal locations for wind farming in the country.

REFERENCES

AZZALINI, A. and CAPITANIO, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 367–389. MR1983753 https://doi.org/10.1111/1467-9868.00391

BHATNAGAR, S., CHANG, W., KIM, S. and WANG, J. (2020). Computer model calibration with time series data using deep learning and quantile regression. Available at arXiv:2008.13066.

BOLTZ, S., DEBREUVE, E. and BARLAUD, M. (2007). kNN-based high-dimensional Kullback–Leibler distance for tracking. In *Eighth International Workshop on Image Analysis for Multimedia Interactive Services* (*WIAMIS* '07) 16–16.

BRITISH PETROLEUM (2020). BP statistical review of world energy. Available at www.bp.com/content/dam/bp/en/corporate/pdf/energy-economics/statistical-review/bp-stats-review-2018-full-report.pdf.

CANNON, A. J. (2018). Multivariate quantile mapping bias correction: An N-dimensional probability density function transform for climate model simulations of multiple variables. *Clim. Dyn.* **50** 31–49.

CANNON, A. J., PIANI, C. and SIPPEL, S. (2020). Chapter 5—bias correction of climate model output for impact models. In *Climate Extremes and Their Implications for Impact and Risk Assessment* (J. Sillmann, S. Sippel and S. Russo, eds.) 77–104. https://doi.org/10.1016/B978-0-12-814895-2.00005-7

CANNON, A. J., SOBIE, S. R. and MURDOCK, T. Q. (2015). Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *J. Climate* **28** 6938–6959. https://doi.org/10.1175/JCLI-D-14-00754.1

CASTRUCCIO, S., OMBAO, H. and GENTON, M. G. (2018). A scalable multi-resolution spatio-temporal model for brain activation and connectivity in fMRI data. *Biometrics* **74** 823–833. MR3860703 https://doi.org/10.1111/biom.12844

CHEN, W., CASTRUCCIO, S. and GENTON, M. G. (2021). Assessing the risk of disruption of wind turbine operations in Saudi Arabia using Bayesian spatial extremes. *Extremes* **24** 267–292. MR4246278 https://doi.org/10.1007/s10687-020-00384-1

Key words and phrases. Bias correction, Kullback–Leibler divergence, non-Gaussian process, nonstationary model, spatiotemporal model, wind energy.

CHEN, L., PRYOR, S. C. and LI, D. (2012). Assessing the performance of intergovernmental panel on climate change AR5 climate models in simulating and projecting wind speeds over China. *J. Geophys. Res., Atmos.* **117** D24102. https://doi.org/10.1029/2012JD017533

CHEN, W., CASTRUCCIO, S., GENTON, M. G. and CRIPPA, P. (2018). Current and future estimates of wind energy potential over Saudi Arabia. *J. Geophys. Res., Atmos.* **123** 6443–6459. https://doi.org/10.1029/2017JD028212

CRIPPA, P., ALIFA, M., BOLSTER, D., GENTON, M. G. and CASTRUCCIO, S. (2021). A heteroskedastic time-varying model for improved hourly wind power forecasting. Under review.

DING, Y. (2019). *Data Science for Wind Energy*. CRC press, Boca Raton, FL.

DUNNE, J. P., JOHN, J. G., SHEVLIAKOVA, E., STOUFFER, R. J., KRASTING, J. P., MALYSHEV, S. L., MILLY, P. C. D., SENTMAN, L. T., ADCROFT, A. J. et al. (2013). GFDL's ESM2 global coupled climate–carbon Earth system models. Part II: Carbon system formulation and baseline simulation characteristics. *J. Climate* **26** 2247–2267. https://doi.org/10.1175/JCLI-D-12-00150.1

EMEIS, S. (2018). *Wind Energy Meteorology*, 2nd ed. Springer, New York. https://doi.org/10.1007/978-3-319-72859-9

ERDIN, R., FREI, C. and KÜNSCH, H. R. (2012). Data transformation and uncertainty in geostatistical combination of radar and rain gauges. *J. Hydrometeorol.* **13** 1332–1346. https://doi.org/10.1175/JHM-D-11-096.1

FARAG, A. A. (2019). The story of NEOM city: Opportunities and challenges. In *New Cities and Community Extensions in Egypt and the Middle East* (A. Ibrahim, S. Attia and Z. Shafik, eds.) 35–49. Springer, Cham, Switzerland.

FRANÇOIS, B., VRAC, M., CANNON, A. J., ROBIN, Y. and ALLARD, D. (2020). Multivariate bias corrections of climate simulations: Which benefits for which losses? *Earth Syst. Dyn.* **11** 537–562. https://doi.org/10.5194/esd-11-537-2020

GELARO, R., MCCARTY, W., SUÁREZ, M. J., TODLING, R., MOLOD, A., TAKACS, L., RANDLES, C. A., DARMENOV, A., BOSILOVICH, M. G. et al. (2017). The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *J. Climate* **30** 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

GIANI, P., FELIPE, T., GENTON, M. G., CASTRUCCIO, S. and CRIPPA, P. (2020). Closing the gap between wind energy targets and implementation for emerging countries. *Appl. Energy* **269** 115085. https://doi.org/10.1016/j.apenergy.2020.115085

GUALTIERI, G. (2019). A comprehensive review on wind resource extrapolation models applied in wind energy. *Renew. Sustain. Energy Rev.* **102** 215–233. https://doi.org/10.1016/j.rser.2018.12.015

HAWKINS, E., OSBORNE, T. M., HO, C. K. and CHALLINOR, A. J. (2013). Calibration and bias correction of climate projections for crop modelling: An idealised case study over Europe. *Agric. For. Meteorol.* **170** 19–31. Agricultural prediction using climate model ensembles. https://doi.org/10.1016/j.agrformet.2012.04.007

HEMER, M., MCINNES, K. and RANASINGHE, R. (2012). Climate and variability bias adjustment of climate model-derived winds for a southeast Australian dynamical wave model. *Ocean Dyn.* **62** 87–104.

HERSBACH, H., BELL, B., BERRISFORD, P., HIRAHARA, S., HORÁNYI, A., MUÑOZ-SABATER, J., NICOLAS, J., PEUBEY, C., RADU, R. et al. (2020). The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146** 1999–2049. https://doi.org/10.1002/qj.3803

HIGDON, D. (2002). Space and space–time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* 37–56. Springer, London. MR2059819

HO, C. K., STEPHENSON, D. B., COLLINS, M., FERRO, C. A. T. and BROWN, S. J. (2012). Calibration strategies: A source of additional uncertainty in climate change projections. *Bull. Am. Meteorol. Soc.* **93** 21–26. https://doi.org/10.1175/2011BAMS3110.1

INTERNATIONAL RENEWABLE ENERGY AGENCY (2018). Renewable energy statistics. Available at https://irena.org/publications/2018/Jul/Renewable-Energy-Statistics-2018.

INTERNATIONAL RENEWABLE ENERGY AGENCY (2019). Renewable energy market analysis: GCC. Available at https://www.irena.org/publications/2019.

IPCC (2014). Part A: Global and sectoral aspects. In *AR5 Climate Change* 2014: *Impacts*, *Adaptation*, *and Vulnerability* Cambridge Univ. Press.

JEONG, J., CASTRUCCIO, S., CRIPPA, P. and GENTON, M. G. (2018). Reducing storage of global wind ensembles with stochastic generators. *Ann. Appl. Stat.* **12** 490–509. MR3773402 https://doi.org/10.1214/17-AOAS1105

JEONG, J., YAN, Y., CASTRUCCIO, S. and GENTON, M. G. (2019). A stochastic generator of global monthly wind energy with Tukey g-and-h autoregressive processes. *Statist. Sinica* **29** 1105–1126. MR3932511

KENNEDY, M. C. and O'HAGAN, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika* **87** 1–13. MR1766824 https://doi.org/10.1093/biomet/87.1.1

KENNEDY, M. C. and O'HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. MR1858398 https://doi.org/10.1111/1467-9868.00294

KIM, K. B., KWON, H.-H. and HAN, D. (2015). Bias correction methods for regional climate model simulations considering the distributional parametric uncertainty underlying the observations. *J. Hydrol.* **530** 568–579. https://doi.org/10.1016/j.jhydrol.2015.10.015

KINLEY, R. (2017). Climate change after Paris: From turning point to transformation. *Climate Policy* **17** 9–15. https://doi.org/10.1080/14693062.2016.1191009

KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* **22** 79–86. MR0039968 https://doi.org/10.1214/aoms/1177729694

LEEDS, W. B., MOYER, E. J. and STEIN, M. L. (2015). Simulation of future climate under changing temporal covariance structures. *Adv. Stat. Climatol. Meteorol. Oceanogr.* **1** 1–14. https://doi.org/10.5194/ascmo-1-1-2015

LI, D., FENG, J., XU, Z., YIN, B., SHI, H. and QI, J. (2019). Statistical bias correction for simulated wind speeds over CORDEX-East Asia. *Earth Space Sci.* **6** 200–211. https://doi.org/10.1029/2018EA000493

MEHROTRA, R. and SHARMA, A. (2016). A multivariate quantile-matching bias correction approach with auto- and cross-dependence across multiple time scales: Implications for downscaling. *J. Climate* **29** 3519–3539.

NGUYEN, H., MEHROTRA, R. and SHARMA, A. (2019). Correcting systematic biases across multiple atmospheric variables in the frequency domain. *Clim. Dyn.* **52** 1283–1298.

NREP (2018). Saudi Arabia renewable energy targets and long term visibility, national renewable energy program.

NURUNNABI, M. (2017). Transformation from an oil-based economy to a knowledge-based economy in Saudi Arabia: The direction of saudi vision 2030. *Journal of the Knowledge Economy* **8** 536–64.

PACIOREK, C. J. and SCHERVISH, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* **17** 483–506. MR2240939 https://doi.org/10.1002/env.785

PALACIOS, M. B. and STEEL, M. F. J. (2006). Non-Gaussian Bayesian geostatistical modeling. *J. Amer. Statist. Assoc.* **101** 604–618. MR2281244 https://doi.org/10.1198/016214505000001195

POPPICK, A., MCINERNEY, D. J., MOYER, E. J. and STEIN, M. L. (2016). Temperatures in transient climates: Improved methods for simulations with evolving temporal covariances. *Ann. Appl. Stat.* **10** 477–505. MR3480504 https://doi.org/10.1214/16-AOAS903

REHMAN, S., EL-AMIN, I. M., AHMAD, F., SHAAHID, S. M., AL-SHEHRI, A. M. and BAKHASHWAIN, J. M. (2007). Wind power resource assessment for Rafha, Saudi Arabia. *Renew. Sustain. Energy Rev.* **11** 937–950. https://doi.org/10.1016/j.rser.2005.07.003

REN21 SECRETARIAT (2018). Renewables 2018—global status report. Paris, France.

RISSER, M. and CALDER, C. (2017). Local likelihood estimation for covariance functions with spatially-varying parameters: The convoSPAT package for R. *J. Stat. Softw.* **81** 1–32. https://doi.org/10.18637/jss.v081.i14

STEIN, M. L. (1999). *Interpolation of Spatial Data*: *Some Theory for Kriging*. *Springer Series in Statistics*. Springer, New York. MR1697409 https://doi.org/10.1007/978-1-4612-1494-6

TAGLE, F., CASTRUCCIO, S. and GENTON, M. G. (2020). A hierarchical bi-resolution spatial skew-*t* model. *Spat. Stat.* **35** 100398, 12. MR4052623 https://doi.org/10.1016/j.spasta.2019.100398

TAGLE, F., CASTRUCCIO, S., CRIPPA, P. and GENTON, M. G. (2019). A non-Gaussian spatio-temporal model for daily wind speeds based on a multi-variate skew-*t* distribution. *J. Time Series Anal.* **40** 312–326. MR3946155 https://doi.org/10.1111/jtsa.12437

TAGLE, F., GENTON, M. G., YIP, A., MOSTAMANDI, S., STENCHIKOV, G. and CASTRUCCIO, S. (2020). A high-resolution bilevel skew-t stochastic generator for assessing Saudi Arabia's wind energy resources. *Environmetrics* **31** e2628, 16. MR4166851 https://doi.org/10.1002/env.2628

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. https://doi.org/10.1175/BAMS-D-11-00094.1

TEUTSCHBEIN, C. and SEIBERT, J. (2012). Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *J. Hydrol.* **456–457** 12–29. https://doi.org/10.1016/j.jhydrol.2012.05.052

TUO, R. and WU, C. F. J. (2015). Efficient calibration for imperfect computer models. *Ann. Statist.* **43** 2331–2352. MR3405596 https://doi.org/10.1214/15-AOS1314

UNFCCC (2020). Intended nationally determined contributions: Kingdom of Saudi Arabia. Available at www4.unfccc.int/sites/submissions/INDC/Published%20Documents/Saudi%20Arabia/1/KSA-INDCs%20English.pdf.

VAN VUUREN, D. P., EDMONDS, J., KAINUMA, M., RIAHI, K., THOMSON, A., HIBBARD, K., HURTT, G. C., KRAM, T., KREY, V. et al. (2011). The representative concentration pathways: An overview. *Clim. Change* **109** 5–31.

VASSALLO, D., KRISHNAMURTHY, R. and FERNANDO, H. J. S. (2020). Decreasing wind speed extrapolation error via domain-specific feature extraction and selection. *Wind Energy Science* **5** 959–975. https://doi.org/10.5194/wes-5-959-2020

WANG, Q., KULKARNI, S. R. and VERDÚ, S. (2009). Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Trans. Inf. Theory* **55** 2392–2405. MR2729888 https://doi.org/10.1109/TIT.2009.2016060

WORLD BANK (2020). Energy use (kg of oil equivalent per capita). Available at https://data.worldbank.org/indicator/EG.USE.PCAP.KG.OE.

YEO, I.-K. and JOHNSON, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika* **87** 954–959. MR1813988 https://doi.org/10.1093/biomet/87.4.954

YUAN, Q., THORARINSDOTTIR, T. L., BELDRING, S., WONG, W. K., HUANG, S. and XU, C.-Y. (2019). New approach for bias correction and stochastic downscaling of future projections for daily mean temperatures to a high-resolution grid. *J. Appl. Meteorol. Climatol.* **58** 2617–2632. https://doi.org/10.1175/JAMC-D-19-0086.1

ZHANG, J., CRIPPA, P., GENTON, M. G. and CASTRUCCIO, S. (2021). Supplement to "Assessing the reliability of wind power operations under a changing climate with a non-Gaussian bias correction." https://doi.org/10.1214/21-AOAS1460SUPP

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1850–1871
https://doi.org/10.1214/21-AOAS1490

NONPARAMETRIC IMPORTANCE SAMPLING FOR WIND TURBINE RELIABILITY ANALYSIS WITH STOCHASTIC COMPUTER MODELS

BY SHUORAN LI[1], YOUNG MYOUNG KO[2] AND EUNSHIN BYON[3]

[1]*Department of Statistics, University of Pittsburgh, SHL198@pitt.edu*
[2]*Department of Industrial and Management Engineering, Pohang University of Science and Technology, youngko@postech.ac.kr*
[3]*Department of Industrial and Operations Engineering, University of Michigan, ebyon@umich.edu*

Using aeroelastic stochastic simulations, this study presents an importance sampling method for assessing wind turbine reliability. As the size of modern wind turbines gets larger, structural reliability analysis becomes more important to prevent any catastrophic failures. At the design stage, operational data do not exist or are scarce. Therefore, aeroelastic simulation is often employed for reliability analysis. Importance sampling is one of the powerful variance reduction techniques to mitigate computational burden in stochastic simulations. In the literature, wind turbine reliability assessment with importance sampling has been studied with a single variable, wind speed. However, other atmospheric stability conditions also impose substantial stress on the turbine structure. Moreover, each environmental factor's effect on the turbine's load response depends on other factors. This study investigates how multiple environmental factors collectively affect the turbine reliability. Specifically, we devise a new nonparametric importance sampling method that can quantify the contributions of each environmental factor and its interactions with other factors, while avoiding computational problems and data sparsity issue arising in rare event simulation. Our wind turbine case study and numerical examples demonstrate the advantage of the proposed approach.

REFERENCES

ACKLEY, D. (2012). *A Connectionist Machine for Genetic Hillclimbing* **28**. Springer, Berlin.

ADVANCE METROLOGY LAB (2020). Inland Wind Farm Dataset2 (WT2). Available at https://aml.engr.tamu.edu/book-dswe/dswe-datasets/.

AZIZ EZZAT, A., JUN, M. and DING, Y. (2019). Spatio-temporal short-term wind forecast: A calibrated regime-switching method. *Ann. Appl. Stat.* **13** 1484–1510. MR4019147 https://doi.org/10.1214/19-AOAS1243

BIERENS, H. J. (1987). Kernel estimators of regression functions. In *Advances in Econometrics*: *Fifth World Congress* **1** 99–144. Cambridge Univ. Press, New York.

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. *Information Science and Statistics*. Springer, New York. MR2247587 https://doi.org/10.1007/978-0-387-45528-0

CANNAMELA, C., GARNIER, J. and IOOSS, B. (2008). Controlled stratification for quantile estimation. *Ann. Appl. Stat.* **2** 1554–1580. MR2655671 https://doi.org/10.1214/08-AOAS186

CAO, Q. D. and CHOE, Y. (2019). Cross-entropy based importance sampling for stochastic simulation models. *Reliab. Eng. Syst. Saf.* **191** 106526.

CHEN, Y.-C. and CHOE, Y. (2019). Importance sampling and its optimality for stochastic simulation models. *Electron. J. Stat.* **13** 3386–3423. MR4010983 https://doi.org/10.1214/19-ejs1604

CHOE, Y., BYON, E. and CHEN, N. (2015). Importance sampling for reliability evaluation with stochastic simulation models. *Technometrics* **57** 351–361. MR3384950 https://doi.org/10.1080/00401706.2014.1001523

CHOE, Y., LAM, H. and BYON, E. (2018). Uncertainty quantification of stochastic simulation for black-box computer experiments. *Methodol. Comput. Appl. Probab.* **20** 1155–1172. MR3873620 https://doi.org/10.1007/s11009-017-9599-7

CHOE, Y., PAN, Q. and BYON, E. (2016). Computationally efficient uncertainty minimization in wind turbine extreme load assessments. *J. Sol. Energy Eng.* **138** 041012.

Key words and phrases. Kernel regression, Monte Carlo simulation, rare event simulation, variance reduction, wind energy.

CHOI, K.-S., HUH, Y.-H., KWON, I.-B. and YOON, D.-J. (2012). A tip deflection calculation method for a wind turbine blade using temperature compensated FBG sensors. *Smart Materials and Structures* **21** 025008.
DE BOER, P.-T., KROESE, D. P., MANNOR, S. and RUBINSTEIN, R. Y. (2005). A tutorial on the cross-entropy method. *Ann. Oper. Res.* **134** 19–67. MR2136658 https://doi.org/10.1007/s10479-005-5724-z
DING, Y. (2019). *Data Science for Wind Energy*. CRC Press, Boca Raton, FL.
DUBOURG, V., SUDRET, B. and DEHEEGER, F. (2013). Metamodel-based importance sampling for structural reliability analysis. *Probab. Eng. Mech.* **33** 47–57.
FAN, J. and GIJBELS, I. (2018). *Local Polynomial Modelling and Its Applications*: *Monographs on Statistics and Applied Probability* **66**. Routledge, London. MR1383587
FAN, J. and YIM, T. H. (2004). A cross-validation method for estimating conditional densities. *Biometrika* **91** 819–834. MR2126035 https://doi.org/10.1093/biomet/91.4.819
GIVENS, G. H. and RAFTERY, A. E. (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *J. Amer. Statist. Assoc.* **91** 132–141. MR1394067 https://doi.org/10.2307/2291389
GUALTIERI, G. (2016). Atmospheric stability varying wind shear coefficients to improve wind resource extrapolation: A temporal analysis. *Renew. Energy* **87** 376–390.
HANSEN, B. E. (2009). Lecture notes on nonparametrics. Available at https://www.ssc.wisc.edu/~bhansen/718/NonParametrics2.pdf.
INTERNATIONAL ELECTROTECHNICAL COMMISSION (2005). Wind Turbines—Part 1: Design Requirements, IEC/TC88,61400-1 ed.3.
JONKMAN, B. J. (2009). TurbSim user's guide: Version 1.50. Golden, CO: National Renewable Energy Laboratory.
JONKMAN, J. M. and BUHL JR, M. L. (2005). FAST user's guide. Golden, CO: National Renewable Energy Laboratory.
JONKMAN, J., BUTTERFIELD, S., MUSIAL, W. and SCOTT, G. (2005). Definition of a 5-MW reference wind turbine for offshore system development. Golden, CO: National Renewable Energy Laboratory.
KO, Y. M. and BYON, E. (2021). Optimal budget allocation for stochastic simulation with importance sampling: Exploration vs. replication. *IISE Trans*. To appear.
KONG, C., BANG, J. and SUGIYAMA, Y. (2005). Structural investigation of composite wind turbine blade considering various load cases and fatigue life. *Energy* **30** 2101–2114.
KURTZ, N. and SONG, J. (2013). Cross-entropy-based adaptive importance sampling using Gaussian mixture. *Struct. Saf.* **42** 35–44.
LEE, G., BYON, E., NTAIMO, L. and DING, Y. (2013). Bayesian spline method for assessing extreme loads on wind turbines. *Ann. Appl. Stat.* **7** 2034–2061. MR3161712 https://doi.org/10.1214/13-AOAS670
LEE, G., DING, Y., GENTON, M. G. and XIE, L. (2015). Power curve estimation with multivariate environmental factors for inland and offshore wind farms. *J. Amer. Statist. Assoc.* **110** 56–67. MR3338486 https://doi.org/10.1080/01621459.2014.977385
LI, Q. and RACINE, J. S. (2007). *Nonparametric Econometrics*: *Theory and Practice*. Princeton Univ. Press, Princeton, NJ. MR2283034
MANUEL, L., NGUYEN, H. H. and BARONE, M. F. (2013). On the use of a large database of simulated wind turbine loads to aid in assessing design standard provisions. In *Proceedings of the* 51*st AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition* **4**.
MORIARTY, P. (2008). Database for validation of design load extrapolation techniques. *Wind Energy*: *An International Journal for Progress and Applications in Wind Power Conversion Technology* **11** 559–576.
NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.
NEDDERMEYER, J. C. (2009). Computationally efficient nonparametric importance sampling. *J. Amer. Statist. Assoc.* **104** 788–802. MR2541595 https://doi.org/10.1198/jasa.2009.0122
OWEN, A. B. (2013). Monte Carlo Theory, Methods and Examples. Book in progress. Online version available at https://statweb.stanford.edu/~owen/mc/.
PAN, Q., KO, Y. M. and BYON, E. (2021). Uncertainty quantification for extreme quantile estimation with stochastic computer models. *IEEE Trans. Reliab.* **70** 134–145.
PAN, Q., BYON, E., KO, Y. M. and LAM, H. (2020). Adaptive importance sampling for extreme quantile estimation with stochastic black box computer models. *Naval Res. Logist.* **67** 524–547.
RUBINSTEIN, R. Y. and KROESE, D. P. (2017). *Simulation and the Monte Carlo Method*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3617204
WASSERMAN, L. (2006). *All of Nonparametric Statistics*. *Springer Texts in Statistics*. Springer, New York. MR2172729
WISER, R., HAND, M., SEEL, J. and PAULOS, B. (2016). The Future of Wind Energy, Part 3: Reducing Wind Energy Costs through Increased Turbine Size: Is the Sky the Limit? Available at https://emp.lbl.gov/news/future-wind-energy-part-3-reducing-wind.

WISER, R., BOLINGER, M., HOEN, B., MILLSTEIN, D., RAND, J., BARBOSE, G., DARGHOUTH, N., GORMAN, W., JEONG, S. et al. (2020). Wind Energy Technology Data Update: 2020 Edition. Technical Report, Lawrence Berkeley National Laboratory, Berkeley, CA.

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1872–1896
https://doi.org/10.1214/21-AOAS1472

ESTIMATING ANIMAL UTILIZATION DISTRIBUTIONS FROM MULTIPLE DATA TYPES: A JOINT SPATIOTEMPORAL POINT PROCESS FRAMEWORK

BY JOE WATSON[1], RUTH JOY[2], DOMINIC TOLLIT[3], SHEILA J. THORNTON[4] AND MARIE AUGER-MÉTHÉ[5]

[1]*Department of Statistics, University of British Columbia, joe.watson@stat.ubc.ca*
[2]*School of Environmental Science, Simon Fraser University and SMRU Consulting, rjoy@sfu.ca*
[3]*SMRU Consulting, djt@smruconsulting.com*
[4]*Fisheries and Oceans Canada, Pacific Science Enterprise Centre, Sheila.Thornton@dfo-mpo.gc.ca*
[5]*Institute for the Oceans & Fisheries and the Department of Statistics, University of British Columbia, auger-methe@stat.ubc.ca*

Models of the spatial distribution of animals provide useful tools to help ecologists quantify species-environment relationships, and they are increasingly being used to help determine the impacts of climate and habitat changes on species. While high-quality survey-style data with known effort are sometimes available, often researchers have multiple datasets of varying quality and type. In particular, collections of sightings made by citizen scientists are becoming increasingly common, with no information typically provided on their observer effort. Many standard modelling approaches ignore observer effort completely which can severely bias estimates of an animal's distribution. Combining sightings data from observers who followed different protocols is challenging. Any differences in observer skill, spatial effort and the detectability of the animals across space all need to be accounted for. To achieve this, we build upon the recent advancements made in integrative species distribution models and present a novel marked spatiotemporal point process framework for estimating the utilization distribution (UD) of the individuals of a highly mobile species. We show that, in certain settings, we can also use the framework to combine the UDs from the sampled individuals to estimate the species' distribution. We combine the empirical results from a simulation study with the implications outlined in a causal directed acyclic graph to identify the necessary assumptions required for our framework to control for observer effort when it is unknown. We then apply our framework to combine multiple datasets collected on the endangered Southern Resident Killer Whales to estimate their monthly effort-corrected space-use.

REFERENCES

BACHL, F. E., LINDGREN, F., BORCHERS, D. L. and ILLIAN, J. B. (2019). inlabru: An R package for Bayesian spatial modelling from ecological survey data. *Methods Ecol. Evol.* **10** 760–766.

BADDELEY, A., RUBAK, E. and TURNER, R. (2015). *Spatial Point Patterns*: *Methodology and Applications with R*. Chapman & Hall/CRC, Boca Raton, FL.

BADDELEY, A. and TURNER, R. (2014). Package 'spatstat'. The Comprehensive R Archive Network.

BEDRIÑANA-ROMANO, L., HUCKE-GAETE, R., VIDDI, F. A., MORALES, J., WILLIAMS, R., ASHE, E., GARCÉS-VARGAS, J., TORRES-FLOREZ, J. P. and RUIZ, J. (2018). Integrating multiple data sources for assessing blue whale abundance and distribution in Chilean northern Patagonia. *Divers. Distrib.* **24** 991–1004.

BIVAND, R. S., PEBESMA, E. and GÓMEZ-RUBIO, V. (2013). *Applied Spatial Data Analysis with R*, 2nd ed. *Use R!* Springer, New York. MR3099410 https://doi.org/10.1007/978-1-4614-7618-4

BIVAND, R. and RUNDEL, C. (2013). rgeos: Interface to geometry engine-open source (GEOS). R Package Version 0.3-2.

Key words and phrases. Species distribution models, preferential sampling, spatiotemporal analysis, citizen science.

BIVAND, R., KEITT, T., ROWLINGSON, B., PEBESMA, E., SUMNER, M., HIJMANS, R., ROUAULT, E. and BIVAND, M. R. (2015). Package 'rgdal'. Bindings for the Geospatial Data Abstraction Library. Available at https://cran.r-project.org/web/packages/rgdal/index.html (accessed on 15 October 2017).

CHAKRABORTY, A., GELFAND, A. E., WILSON, A. M., LATIMER, A. M. and SILANDER, J. A. (2011). Point pattern modelling for degraded presence-only data over large regions. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **60** 757–776. MR2844854 https://doi.org/10.1111/j.1467-9876.2011.00769.x

CLARE, J., MCKINNEY, S. T., DEPUE, J. E. and LOFTIN, C. S. (2017). Pairing field methods to improve inference in wildlife surveys while accommodating detection covariance. *Ecol. Appl.* **27** 2031–2047. https://doi.org/10.1002/eap.1587

DFO. Killer Whale (Northeast Pacific Southern Resident Population). http://www.dfo-mpo.gc.ca/species-especes/profiles-profils/killerWhalesouth-PAC-NE-epaulardsud-eng.html. Accessed: 2019-03-29.

DIGGLE, P. J., MENEZES, R. and SU, T. (2010). Geostatistical inference under preferential sampling. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59** 191–232. MR2744471 https://doi.org/10.1111/j.1467-9876.2009.00701.x

DORAZIO, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Glob. Ecol. Biogeogr.* **23** 1472–1484.

ELITH, J. and LEATHWICK, J. (2007). Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Divers. Distrib.* **13** 265–275.

FIEBERG, J. and BÖRGER, L. (2012). Could you please phrase "home range" as a question? *J. Mammal.* **93** 890–902.

FIEBERG, J., SIGNER, J., SMITH, B. and AVGAR, T. (2021). A 'How to' guide for interpreting parameters in habitat-selection analyses. *J. Anim. Ecol.* **90** 1027–1043. https://doi.org/10.1111/1365-2656.13441

FITHIAN, W. and HASTIE, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *Ann. Appl. Stat.* **7** 1917–1939. MR3161707 https://doi.org/10.1214/13-AOAS667

FITHIAN, W., ELITH, J., HASTIE, T. and KEITH, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods Ecol. Evol.* **6** 424–438. https://doi.org/10.1111/2041-210X.12242

FLEMING, C. H., CALABRESE, J. M., MUELLER, T., OLSON, K. A., LEIMGRUBER, P. and FAGAN, W. F. (2014). From fine-scale foraging to home ranges: A semivariance approach to identifying movement modes across spatiotemporal scales. *Amer. Nat.* **183** E154–E167.

FLEMING, C. H., FAGAN, W. F., MUELLER, T., OLSON, K. A., LEIMGRUBER, P. and CALABRESE, J. M. (2015). Rigorous home range estimation with movement data: A new autocorrelated kernel density estimator. *Ecology* **96** 1182–1188.

FORD, J. K., ELLIS, G. M. and BALCOMB, K. C. (1996). *Killer Whales*: *The Natural History and Genealogy of Orcinus Orca in British Columbia and Washington*. UBC Press, Vancouver.

FORD, J. K., PILKINGTON, J. F., OTSUKI, M., GISBORNE, B., ABERNETHY, R., STREDULINSKY, E., TOWERS, J. and ELLIS, G. (2017). *Habitats of Special Importance to Resident Killer Whales* (*Orcinus Orca*) *Off the West Coast of Canada*. Fisheries and Oceans Canada, Ecosystems and Oceans Science.

GELFAND, A. E. (2020). Statistical challenges in spatial analysis of plant ecology data. *Spat. Stat.* **37** 100418, 25. MR4109601 https://doi.org/10.1016/j.spasta.2020.100418

GELFAND, A. E. and SHIROTA, S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecol. Monogr.* **89** e01372.

GELMAN, A., HWANG, J. and VEHTARI, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24** 997–1016. MR3253850 https://doi.org/10.1007/s11222-013-9416-2

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. With comments and a rejoinder by the authors. MR1422404

GIRAUD, C., CALENGE, C., CORON, C. and JULLIARD, R. (2016). Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics* **72** 649–658. MR3515791 https://doi.org/10.1111/biom.12431

GLENNIE, R., BUCKLAND, S. T. and THOMAS, L. (2015). The effect of animal movement on line transect estimates of abundance. *PLoS ONE* **10** e0121333.

GLENNIE, R., BUCKLAND, S. T., LANGROCK, R., GERRODETTE, T., BALLANCE, L. T., CHIVERS, S. J. and SCOTT, M. D. (2021). Incorporating Animal Movement Into Distance Sampling. *J. Amer. Statist. Assoc.* **116** 107–115. MR4227678 https://doi.org/10.1080/01621459.2020.1764362

HAUSER, D. D., VAN BLARICOM, G. R., HOLMES, E. E. and OSBORNE, R. W. (2006). Evaluating the use of whalewatch data in determining killer whale (Orcinus orca) distribution patterns. *J. Cetacean Res. Manag.* **8** 273.

HAUSER, D. D., LOGSDON, M. G., HOLMES, E. E., VANBLARICOM, G. R. and OSBORNE, R. W. (2007). Summer distribution patterns of southern resident killer whales Orcinus orca: Core areas and spatial segregation of social groups. *Mar. Ecol. Prog. Ser.* **351** 301–310.

HEFLEY, T. J. and HOOTEN, M. B. (2016). Hierarchical species distribution models. *Current Landscape Ecology Reports* **1** 87–97.

HERNAN, M. A. and ROBINS, J. M. (2020). *Causal Inference*: *What If*. CRC Press, Boca Raton, FL.

HUSSEY, N. E., KESSEL, S. T., AARESTRUP, K., COOKE, S. J., COWLEY, P. D., FISK, A. T., HARCOURT, R. G., HOLLAND, K. N., IVERSON, S. J. et al. (2015). Aquatic animal telemetry: A panoramic window into the underwater world. *Science* **348**.

JOHNSON, D. S., HOOTEN, M. B. and KUHN, C. E. (2013). Estimating animal resource selection from telemetry data using point process models. *J. Anim. Ecol.* **82** 1155–1164.

JOHNSON, D. S. and LONDON, J. M. (2018). crawl: An R package for fitting continuous-cime correlated random walk models to animal movement data.

JOHNSON, D. S., LONDON, J. M., LEA, M.-A. and DURBAN, J. W. (2008). Continuous-time correlated random walk model for animal telemetry data. *Ecology* **89** 1208–1215.

KOSHKINA, V., WANG, Y., GORDON, A., DORAZIO, R. M., WHITE, M. and STONE, L. (2017). Integrated species distribution models: Combining presence-background data and site-occupancy data with imperfect detection. *Methods Ecol. Evol.* **8** 420–430.

LELE, S. R., MERRILL, E. H., KEIM, J. and BOYCE, M. S. (2013). Selection, use, choice and occupancy: Clarifying concepts in resource selection studies. *J. Anim. Ecol.* **82** 1183–1191.

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. With discussion and a reply by the authors. MR2853727 https://doi.org/10.1111/j.1467-9868.2011.00777.x

LINDGREN, F., RUE, H. et al. (2015). Bayesian spatial modelling with R-INLA. *J. Stat. Softw.* **63** 1–25.

MATTHIOPOULOS, J., FIEBERG, J. and AARTS, G. (2020). *Species-Habitat Associations*: *Spatial Data*, *Predictive Models*, *and Ecological Insights*. Univ. Minnesota Libraries Publishing. Retrieved from the Univ. Minnesota Digital Conservancy, http://hdl.handle.net/11299/217469.

MILLER, D. A., PACIFICI, K., SANDERLIN, J. S. and REICH, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods Ecol. Evol.* **10** 22–37.

MORDECAI, R. S., MATTSSON, B. J., TZILKOWSKI, C. J. and COOPER, R. J. (2011). Addressing challenges when studying mobile or episodic species: Hierarchical Bayes estimation of occupancy and use. *J. Appl. Ecol.* **48** 56–66.

NOAA. Endangered species act status of puget sound killer whales. https://www.westcoast.fisheries.noaa.gov/protected_species/marine_mammals/killer_whale/esa_status.html. Accessed: 2019-09-27.

OLSON, J. K., WOOD, J., OSBORNE, R. W., BARRETT-LENNARD, L. and LARSON, S. (2018). Sightings of southern resident killer whales in the Salish Sea 1976–2014: The importance of a long-term opportunistic dataset. *Endanger. Species Res.* **37** 105–118.

PACIFICI, K., REICH, B. J., MILLER, D. A. W., GARDNER, B., STAUFFER, G., SINGH, S., MCKERROW, A. and COLLAZO, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology* **98** 840–850. https://doi.org/10.1002/ecy.1710

PEBESMA, E. J. and BIVAND, R. S. (2005). Classes and methods for spatial data in R. *R News* **5** 9–13.

PENNINO, M. G., PARADINAS, I., ILLIAN, J. B., MUÑOZ, F., BELLIDO, J. M., LÓPEZ-QUÍLEZ, A. and CONESA, D. (2019). Accounting for preferential sampling in species distribution models. *Ecol. Evol.* **9** 653–663. https://doi.org/10.1002/ece3.4789

R CORE TEAM (2019). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RECHSTEINER, E. U., BIRDSALL, C. F. C., SANDILANDS, D., SMITH, I. U., PHILLIPS, A. V. and BARRETT-LENNARD, L. G. (2013). Quantifying observer effort for opportunistically-collected wildlife sightings. Available at https://killerwhale.vanaqua.org/document.doc?id=140. Accessed: 2019-03-29.

RENNER, I. W. and WARTON, D. I. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics* **69** 274–281. MR3058074 https://doi.org/10.1111/j.1541-0420.2012.01824.x

RENNER, I. W., ELITH, J., BADDELEY, A., FITHIAN, W., HASTIE, T., PHILLIPS, S. J., POPOVIC, G. and WARTON, D. I. (2015). Point process models for presence-only analysis. *Methods Ecol. Evol.* **6** 366–379.

ROYLE, J. A., KERY, M. and GUELAT, J. (2011). Spatial capture–recapture models for search–encounter data. *Methods Ecol. Evol.* **2** 602–611.

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

SEELY, E., OSBORNE, R. W., KOSKI, K. and LARSON, S. (2017). Soundwatch: Eighteen years of monitoring whale watch vessel activities in the Salish Sea. *PLoS ONE* **12** e0189764. https://doi.org/10.1371/journal.pone.0189764

SIGNER, J., FIEBERG, J. and AVGAR, T. (2017). Estimating utilization distributions from fitted step-selection functions. *Ecosphere* **8** e01771.

SIMONS, R. A. (2019). ERDDAP. Available at https://coastwatch.pfeg.noaa.gov/erddap.

SIMPSON, D., ILLIAN, J. B., LINDGREN, F., SØRBYE, S. H. and RUE, H. (2016). Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika* **103** 49–70. MR3465821 https://doi.org/10.1093/biomet/asv064

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. MR1979380 https://doi.org/10.1111/1467-9868.00353

VAN LIESHOUT, M. N. M. (2019). *Theory of Spatial Statistics*: *A Concise Introduction*. CRC Press, Boca Raton, FL.

VANCOUVER AQUARIUM. BC Cetacean Sightings Network. Available at http://wildwhales.org/. Accessed: 2019-03-29.

WARTON, D. I. and SHEPHERD, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Ann. Appl. Stat.* **4** 1383–1402. MR2758333 https://doi.org/10.1214/10-AOAS331

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

WATSON, J., ZIDEK, J. V. and SHADDICK, G. (2019). A general theory for preferential sampling in environmental networks. *Ann. Appl. Stat.* **13** 2662–2700. MR4037445 https://doi.org/10.1214/19-aoas1288

WATSON, J., JOY, R., TOLLIT, D., THORNTON, S. J. and AUGER-MÉTHÉ, M. (2021). Supplement to "Estimating animal utilization distributions from multiple data types: a joint spatiotemporal point process framework." https://doi.org/10.1214/21-AOAS1472SUPP

WHORISKEY, K., MARTINS, E. G., AUGER-MÉTHÉ, M., GUTOWSKY, L. F., LENNOX, R. J., COOKE, S. J., POWER, M. and MILLS FLEMMING, J. (2019). Current and emerging statistical techniques for aquatic telemetry data: A guide to analysing spatially discrete animal detections. *Methods Ecol. Evol.* **10** 935–948.

WOOD, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 3–36. MR2797734 https://doi.org/10.1111/j.1467-9868.2010.00749.x

WORTON, B. J. (1989). Kernel methods for estimating the utilization distribution in home-range studies. *Ecology* **70** 164–168.

YUAN, Y., BACHL, F. E., LINDGREN, F., BORCHERS, D. L., ILLIAN, J. B., BUCKLAND, S. T., RUE, H. and GERRODETTE, T. (2017). Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *Ann. Appl. Stat.* **11** 2270–2297. MR3743297 https://doi.org/10.1214/17-AOAS1078

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1897–1922
https://doi.org/10.1214/21-AOAS1454

SPATIAL VOTING MODELS IN CIRCULAR SPACES: A CASE STUDY OF THE U.S. HOUSE OF REPRESENTATIVES

BY XINGCHEN YU[1] AND ABEL RODRÍGUEZ[2]

[1]*Department of Statistics, University of California, Santa Cruz, xyu26@ucsc.edu*
[2]*Department of Statistics, University of Washington, abelrod@uw.edu*

The use of spatial models for inferring members' preferences from voting data has become widespread in the study of deliberative bodies, such as legislatures. Most established spatial voting models assume that ideal points belong to a Euclidean policy space. However, the geometry of Euclidean spaces (even multidimensional ones) cannot fully accommodate situations in which members at the opposite ends of the ideological spectrum reveal similar preferences by voting together against the rest of the legislature. This kind of voting behavior can arise, for example, when extreme conservatives oppose a measure because they see it as being too costly, while extreme liberals oppose it for not going far enough for them. This paper introduces a new class of spatial voting models in which preferences live in a circular policy space. Such geometry for the latent space is motivated by both theoretical (the so-called "horseshoe theory" of political thinking) and empirical (goodness of fit) considerations. Furthermore, the circular model is flexible and can approximate the one-dimensional version of the Euclidean voting model when the data supports it. We apply our circular model to roll-call voting data from the U.S. Congress between 1988 and 2019 and demonstrate that, starting with the 112th House of Representatives, circular policy spaces consistently provide a better explanation of legislators's behavior than Euclidean ones and that legislators's rankings, generated through the use of the circular geometry, tend to be more consistent with those implied by their stated policy positions.

REFERENCES

ARCENEAUX, K. and NICHOLSON, S. P. (2012). Who wants to have a tea party? The who, what, and why of the Tea Party movement. *PS Polit. Sci. Polit.* **45** 700–710.

BAFUMI, J., GELMAN, A., PARK, D. K. and KAPLAN, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Polit. Anal.* **13** 171–87.

BELKIN, M. and NIYOGI, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems* 585–591.

BESKOS, A., PILLAI, N., ROBERTS, G., SANZ-SERNA, J.-M. and STUART, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli* **19** 1501–1534. MR3129023 https://doi.org/10.3150/12-BEJ414

BETANCOURT, M. J., BYRNE, S. and GIROLAMI, M. (2014). Optimizing the integrator step size for Hamiltonian Monte Carlo. Preprint. Available at arXiv:1411.6669.

BYRNE, S. and GIROLAMI, M. (2013). Geodesic Monte Carlo on embedded manifolds. *Scand. J. Stat.* **40** 825–845. MR3145120 https://doi.org/10.1111/sjos.12036

CARROLL, R., LEWIS, J. B., LO, J., POOLE, K. T. and ROSENTHAL, H. (2013). The structure of utility in spatial models of voting. *Amer. J. Polit. Sci.* **57** 1008–1028.

CLINTON, J. D. and JACKMAN, S. (2009). To simulate or NOMINATE? *Legis. Stud. Q.* **34** 593–621.

CLINTON, J., JACKMAN, S. and RIVERS, D. (2004). The statistical analysis of roll call data. *Am. Polit. Sci. Rev.* **98** 355–370.

CRANE, H. (2017). A hidden Markov model for latent temporal clustering with application to ideological alignment in the U.S. Supreme Court. *Comput. Statist. Data Anal.* **110** 19–36. MR3612605 https://doi.org/10.1016/j.csda.2016.12.010

DAVIS, O. A., HINICH, M. J. and ORDESHOOK, P. C. (1970). An expository development of a mathematical model of the electoral process. *Am. Polit. Sci. Rev.* **64** 426–448.

Key words and phrases. Spatial voting model, roll-call votes, U.S. Congress, factor model, spherical geometry, geodesic Hamiltonian Monte Carlo.

DUCK-MAYR, J. and MONTGOMERY, J. M. (2020). Ends against the middle: Scaling votes when ideological opposites behave the same for antithetical reasons. Technical report, Department of Political Science, Washington Univ. in St. Louis. Available at https://polmeth.org/publications/ends-against-middle-scaling-votes-when-ideological-opposites-behave-same.

EGUIA, J. X. (2013). Challenges to the standard Euclidean spatial model. In *Advances in Political Economy* 169–180. Springer, Berlin.

ENELOW, J. M. and HINICH, M. J. (1984). *The Spatial Theory of Voting*: *An Introduction*. CUP Archive, Cambridge.

FOX, J.-P. (2010). *Bayesian Item Response Modeling*: *Theory and Applications*. *Statistics for Social and Behavioral Sciences*. Springer, New York. MR2657265 https://doi.org/10.1007/978-1-4419-0742-4

GELMAN, A., HWANG, J. and VEHTARI, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat*. *Comput*. **24** 997–1016. MR3253850 https://doi.org/10.1007/s11222-013-9416-2

GELMAN, A. and RUBIN, D. B. (1992). Inferences from iterative simulation using multiple sequences. *Statist*. *Sci*. **7** 457–472.

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

GEWEKE, J. F. and SINGLETON, K. J. (1981). Maximum likelihood "confirmatory" factor analysis of economic time series. *Internat*. *Econom*. *Rev*. **22** 37–54. MR0614346 https://doi.org/10.2307/2526134

GUIMERÀ, R. and SALES-PARDO, M. (2011). Justice blocks and predictability of US Supreme Court votes. *PLoS ONE* **6** e27188.

HARE, C. and POOLE, K. T. (2014). The polarization of contemporary American politics. *Polity* **46** 411–429.

HECKMAN, J. J. and SNYDER JR., J. M. (1996). Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. Technical report, National bureau of economic research.

HOFFMAN, M. D. and GELMAN, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J*. *Mach*. *Learn*. *Res*. **15** 1593–1623. MR3214779

HUMPHREYS, M. and LAVER, M. (2010). Spatial models, cognitive metrics, and majority rule equilibria. *Br*. *J*. *Polit*. *Sci*. **40** 11–30.

JACKMAN, S. (2001). Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference, and model checking. *Polit*. *Anal*. **9** 227–241.

JESSEE, S. A. (2012). *Ideology and Spatial Voting in American Elections*. Cambridge Univ. Press, Cambridge.

KARPOWITZ, C. F., MONSON, J. Q., PATTERSON, K. D. and POPE, J. C. (2011). Tea time in America? The impact of the Tea Party movement on the 2010 midterm elections. *PS Polit*. *Sci*. *Polit*. **44** 303–309.

KINGMA, D. P., WELLING, M. et al. (2019). An introduction to variational autoencoders. *Found*. *Trends Mach*. *Learn*. **12** 307–392.

KRAMER, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE J*. **37** 233–243.

LAWRENCE, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *J*. *Mach*. *Learn*. *Res*. **6** 1783–1816. MR2249872

LEWIS, J. (2019a). Why are Ocasio-Cortez, Omar, Pressley, and Talib estimated to be moderates by NOMINATE? Available at https://voteview.com/articles/Ocasio-Cortez_Omar_Pressley_Tlaib.

LEWIS, J. (2019b). Why is Alexandria Ocasio-Cortez estimated to be a moderate by NOMINATE? Available at https://voteview.com/articles/ocasio_cortez.

LOFLAND, C. L., RODRÍGUEZ, A. and MOSER, S. (2017). Assessing differences in legislators' revealed preferences: A case study on the 107th U.S. Senate. *Ann*. *Appl*. *Stat*. **11** 456–479. MR3634331 https://doi.org/10.1214/16-AOAS951

MARTIN, A. D. and QUINN, K. M. (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999. *Polit*. *Anal*. **10** 134–153.

MCCARTY, N., POOLE, K. T. and ROSENTHAL, H. (2016). *Polarized America*: *The Dance of Ideology and Unequal Riches*, 2nd ed. MIT Press, Cambridge.

MCCORMICK, T. H. and ZHENG, T. (2015). Latent surface models for networks using aggregated relational data. *J*. *Amer*. *Statist*. *Assoc*. **110** 1684–1695. MR3449064 https://doi.org/10.1080/01621459.2014.991395

MCFADDEN, D. (1973). Conditional logit analysis of qualitative choice behavior. In *Frontiers of Economics* (P. Zarembka, ed.) 105–142. Institute of Urban and Regional Development, Univ. California.

MOSER, S., RODRIGUEZ, A. and LOFLAND, C. L. (2021). Multiple ideal points: Revealed preferences in different domains. *Polit*. *Anal*. **29** 139–166.

PIERRE, F. J. (2002). *Le siècle des idéologies*. Pocket, Paris.

POOLE, K. T. and ROSENTHAL, H. (1985). A spatial model for legislative roll call analysis. *Amer*. *J*. *Polit*. *Sci*. **29** 357–384.

POOLE, K. T. and ROSENTHAL, H. (1991). Patterns of congressional voting. *Amer. J. Polit. Sci.* **35** 228–278.
POOLE, K. T. and ROSENTHAL, H. (1997). *Congress: A Political-Economic History of Roll Call Voting*. Oxford Univ. Press, London.
RAGUSA, J. M. and GASPAR, A. (2016). Where's the Tea Party? An examination of the Tea Party's voting behavior in the House of Representatives. *Polit. Res. Q.* **69** 361–372.
RODRÍGUEZ, A. and MOSER, S. (2015). Measuring and accounting for strategic abstentions in the US Senate, 1989–2012. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 779–797. MR3415950 https://doi.org/10.1111/rssc.12099
ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326.
SKOCPOL, T. and WILLIAMSON, V. (2016). *The Tea Party and the Remaking of Republican Conservatism*. Oxford Univ. Press, London.
SMITH, A. L., ASTA, D. M. and CALDER, C. A. (2019). The geometry of continuous latent space models for network data. *Statist. Sci.* **34** 428–453. MR4017522 https://doi.org/10.1214/19-STS702
SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2014). The deviance information criterion: 12 years on. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 485–493. MR3210727 https://doi.org/10.1111/rssb.12062
SPIRLING, A. and MCLEAN, I. (2007). UK OC OK? Interpreting optimal classification scores for the UK House of Commons. *Polit. Anal.* **15** 85–96.
SPIRLING, A. and QUINN, K. (2010). Identifying intraparty voting blocs in the U.K. House of Commons. *J. Amer. Statist. Assoc.* **105** 447–457. MR2724838 https://doi.org/10.1198/jasa.2009.ap07115
TAYLOR, J. (2006). *Where Did the Party Go?: William Jennings Bryan, Hubert Humphrey, and the Jeffersonian Legacy*. Univ. Missouri Press, Columbia.
TENENBAUM, J. B., DE SILVA, V. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.
VAN DER LINDEN, W. J. and HAMBLETON, R. K., eds. (1997). *Handbook of Modern Item Response Theory* Springer, New York. MR1601043 https://doi.org/10.1007/978-1-4757-2691-6
WATANABE, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11** 3571–3594. MR2756194
WATANABE, S. (2013). A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14** 867–897. MR3049492
WEISBERG, H. F. (1974). Dimensionland: An excursion into spaces. *Amer. J. Polit. Sci.* **18** 743–776.
WILSON, M. and DE BOECK, P. (2004). Descriptive and explanatory item response models. In *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. Stat. Soc. Sci. Public Policy* 43–74. Springer, New York. MR2083195 https://doi.org/10.1007/978-1-4757-3990-9_2
YU, X. and RODRIGUEZ, A. (2021a). Supplement to "Spatial voting models in circular spaces: A case study of the U.S. House of Representatives." https://doi.org/10.1214/21-AOAS1454SUPPA
YU, X. and RODRIGUEZ, A. (2021b). Supplement to source code "Spatial voting models in circular spaces: A case study of the U.S. House of Representatives." https://doi.org/10.1214/21-AOAS1454SUPPB
ZHANG, Z. and ZHA, H. (2004). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Sci. Comput.* **26** 313–338. MR2114346 https://doi.org/10.1137/S1064827502419154

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1923–1944
https://doi.org/10.1214/21-AOAS1483

MODELING THE SOCIAL MEDIA RELATIONSHIPS OF IRISH POLITICIANS USING A GENERALIZED LATENT SPACE STOCHASTIC BLOCKMODEL

BY TIN LOK JAMES NG[1], THOMAS BRENDAN MURPHY[2], TED WESTLING[3], TYLER H. MCCORMICK[4] AND BAILEY FOSDICK[5]

[1]*School of Computer Science and Statistics, Trinity College Dublin, ngja@tcd.ie*
[2]*School of Mathematics and Statistics, University College Dublin, brendan.murphy@ucd.ie*
[3]*Department of Mathematics and Statistics, University of Massachusetts Amherst, twestling@umass.edu*
[4]*Department of Statistics and Department of Sociology, University of Washington, tylermc@u.washington.edu*
[5]*Department of Statistics, Colorado State University, bailey.fosdick@colostate.edu*

Dáil Éireann is the principal chamber of the Irish parliament. The 31st Dáil was in session from March 11th, 2011 to February 6th, 2016. Many of the members of the Dáil were active on social media, and many were Twitter users who followed other members of the Dáil. The pattern of Twitter following amongst these politicians provides insights into political alignment within the Dáil. We propose a new model, called the *generalized latent space stochastic blockmodel*, which extends and generalizes both the latent space model and the stochastic blockmodel to study social media connections between members of the Dáil. The probability of an edge between two nodes in a network depends on their respective class labels, as well as sender and receiver effects and latent positions in an unobserved latent space. The proposed model is capable of representing transitivity and clustering, as well as disassortative mixing. A Bayesian method with Markov chain Monte Carlo sampling is proposed for estimation of model parameters. Model selection is performed using the WAIC criterion and models of different number of classes or dimensions of latent space are compared. We use the model to study Twitter following relationships of members of the Dáil and interpret structure found in these relationships. We find that the following relationships amongst politicians is mainly driven by past and present political party membership. We also find that the modeling outputs are informative when studying voting within the Dáil.

REFERENCES

AIROLDI, E. M. (2006). *Bayesian Mixed-Membership Models of Complex and Evolving Networks*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.), Carnegie Mellon Univ. MR2709791

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. *Information Science and Statistics*. Springer, New York. MR2247587 https://doi.org/10.1007/978-0-387-45528-0

BOLLEYER, N. and WEEKS, L. (2009). The puzzle of non-party actors in party democracy: Independents in Ireland. *Comp. Eur. Polit.* **7** 299–324.

CELEUX, G., HURN, M. and ROBERT, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95** 957–970. MR1804450 https://doi.org/10.2307/2669477

COAKLEY, J. and GALLAGHER, M. (2018). *Politics of the Republic of Ireland*, 6th ed. Routledge, London.

CÔME, E. and LATOUCHE, P. (2015). Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Stat. Model.* **15** 564–589. MR3441229 https://doi.org/10.1177/1471082X15577017

FARRELL, D. M., MAIR, P., Ó MUINEACHÁIN S and WALL, M. (2015). Courting but not always serving: Perverted burkeanism and the puzzle of Irish parliamentary cohesion. In *Parties and Party Systems*: *Structure and Context* (R. Johnson and C. Sharman, eds.) 92–107. UBC Press, Vancouver.

Key words and phrases. Social network analysis, latent space stochastic blockmodel, Irish politician.

FOSDICK, B. K., MCCORMICK, T. H., MURPHY, T. B., NG, T. L. J. and WESTLING, T. (2019). Multiresolution network models. *J. Comput. Graph. Statist.* **28** 185–196. MR3939381 https://doi.org/10.1080/10618600.2018.1505633

GELMAN, A., HWANG, J. and VEHTARI, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24** 997–1016. MR3253850 https://doi.org/10.1007/s11222-013-9416-2

GORMLEY, I. C. and MURPHY, T. B. (2008a). Exploring voting blocs within the Irish electorate: A mixture modeling approach. *J. Amer. Statist. Assoc.* **103** 1014–1027. MR2528824 https://doi.org/10.1198/016214507000001049

GORMLEY, I. C. and MURPHY, T. B. (2008b). A mixture of experts model for rank data with applications in election studies. *Ann. Appl. Stat.* **2** 1452–1477. MR2655667 https://doi.org/10.1214/08-AOAS178

HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354. MR2364300 https://doi.org/10.1111/j.1467-985X.2007.00471.x

HANSEN, M. E. (2009). The positions of Irish parliamentary parties 1937–2006. *Ir. Polit. Stud.* **24** 29–44.

HANSEN, M. E. (2010). The parliamentary behaviour of minor parties and independents in Dáil Éireann. *Ir. Polit. Stud.* **25** 643–660.

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

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

KRIVITSKY, P. N., HANDCOCK, M. S., RAFTERY, A. E. and HOFF, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Soc. Netw.* **31** 204–213. https://doi.org/10.1016/j.socnet.2009.04.001

LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2011). Overlapping stochastic block models with application to the French political blogosphere. *Ann. Appl. Stat.* **5** 309–336. MR2810399 https://doi.org/10.1214/10-AOAS382

LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2014). Model selection in overlapping stochastic block models. *Electron. J. Stat.* **8** 762–794. MR3217788 https://doi.org/10.1214/14-EJS903

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equations of state calculations by fast computing machine. *J. Chem. Phys.* **21** 1087–1091.

NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96** 1077–1087. MR1947255 https://doi.org/10.1198/016214501753208735

PENG, L. and CARVALHO, L. (2016). Bayesian degree-corrected stochastic blockmodels for community detection. *Electron. J. Stat.* **10** 2746–2779. MR3549018 https://doi.org/10.1214/16-EJS1163

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712 https://doi.org/10.1080/01621459.2013.829001

RYAN, C., WYSE, J. and FRIEL, N. (2017). Bayesian model selection for the latent position cluster model for social networks. *Netw. Sci.* **5** 70–91.

SALTER-TOWNSHEND, M. and MURPHY, T. B. (2013). Variational Bayesian inference for the latent position cluster model for network data. *Comput. Statist. Data Anal.* **57** 661–671. MR2981116 https://doi.org/10.1016/j.csda.2012.08.004

SALTER-TOWNSHEND, M., WHITE, A., GOLLINI, I. and MURPHY, T. B. (2012). Review of statistical network analysis: Models, algorithms, and software. *Stat. Anal. Data Min.* **5** 260–264. MR2958152 https://doi.org/10.1002/sam.11146

SCHWEINBERGER, M. and HANDCOCK, M. S. (2015). Local dependence in random graph models: Characterization, properties and statistical inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 647–676. MR3351449 https://doi.org/10.1111/rssb.12081

SIBSON, R. (1979). Studies in the robustness of multidimensional scaling: Perturbational analysis of classical scaling. *J. Roy. Statist. Soc. Ser. B* **41** 217–229. MR0547248

SMÍDL, V. and QUINN, A. (2006). *The Variational Bayes Method in Signal Processing*. Springer, Berlin.

SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14** 75–100. MR1449742 https://doi.org/10.1007/s003579900004

WANG, Y. X. R. and BICKEL, P. J. (2017). Likelihood-based model selection for stochastic block models. *Ann. Statist.* **45** 500–528. MR3650391 https://doi.org/10.1214/16-AOS1457

WANG, Y. J. and WONG, G. Y. (1987). Stochastic blockmodels for directed graphs. *J. Amer. Statist. Assoc.* **82** 8–19. MR0883333

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

WEEKS, L. (2009). We don't like (to) party. A typology of independents in Irish political life, 1922–2007. *Ir. Polit. Stud.* **24** 1–27.

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1945–1958
https://doi.org/10.1214/21-AOAS1470

PAN-DISEASE CLUSTERING ANALYSIS OF THE TREND OF PERIOD PREVALENCE

BY SNEHA JADHAV[1], CHENJIN MA[2], YEFEI JIANG[3,*], BEN-CHANG SHIA[3,†] AND SHUANGGE MA[4]

[1]*Department of Mathematics and Statistics, Wake Forest University, jadhavs@wfu.edu*
[2]*College of Statistics and Data Science, Faculty of Science, Beijing University of Technology, mayuan1212@126.com*
[3]*College of Management, Fu Jen Catholic University, *ffay10@gmail.com; †025674@mail.fju.edu.tw*
[4]*Department of Biostatistics, Yale School of Public Health, shuangge.ma@yale.edu*

Prevalence is of essential importance in biomedical and public health research. In the "classic" paradigm it has been studied for each disease individually. Accumulating evidence has shown that diseases can be "correlated." Joint analysis of prevalence can potentially provide important insights beyond individual-disease analysis but has not been well pursued. In this study we take advantage of the unique Taiwan National Health Insurance Research Database (NHIRD) and conduct the first pan-disease analysis of period prevalence trend. The goal is to identify clusters within which diseases have similar period prevalence trends. A novel penalization pursuit approach is applied which has an intuitive formulation and preferable numerical performance. In data analysis the period prevalence values are computed using the records on close to one million subjects and 14 years of observation. With 405 diseases, 35 clusters with sizes larger than one and 27 clusters with sizes one are identified. The clustering results have sound interpretations and differ significantly from those of the alternatives.

REFERENCES

ABDOLLAH, F., GANDAGLIA, G., THURET, R., SCHMITGES, J., TIAN, Z., JELDRES, C., PASSONI, N. M., BRIGANTI, A., SHARIAT, S. F. et al. (2013). Incidence, survival and mortality rates of stage-specific bladder cancer in United States: A trend analysis. *Cancer Epidemiol*. **37** 219–225.

ABRAHAM, C., CORNILLON, P. A., MATZNER-LØBER, E. and MOLINARI, N. (2003). Unsupervised curve clustering using B-splines. *Scand. J. Stat*. **30** 581–595. MR2002229 https://doi.org/10.1111/1467-9469.00350

ANNEGERS, J. F., HAUSER, W. A., BEGHI, E., NICOLOSI, A. and KURLAND, L. T. (1988). The risk of unprovoked seizures after encephalitis and meningitis. *Neurology* **38** 1407–1410. https://doi.org/10.1212/wnl.38.9.1407

AUSTIN, M. A., HOKANSON, J. E. and EDWARDS, K. L. (1998). Hypertriglyceridemia as a cardiovascular risk factor. *Am. J. Cardiol*. **81** 7B–12B.

BONNEAU, C., PERRIN, M., KOSKAS, M., GENIN, A. S. and ROUZIER, R. (2014). Epidemiology and risk factors for cancer of the uterus. *La Revue du Praticien*. **64** 774–779.

BOUVEYRON, C., CÔME, E. and JACQUES, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *Ann. Appl. Stat*. **9** 1726–1760. MR3456352 https://doi.org/10.1214/15-AOAS861

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

CANTO, J. G., SHLIPAK, M. G., ROGERS, W. J., MALMGREN, J. A., FREDERICK, P. D., LAMBREW, C. T., ORNATO, J. P., BARRON, H. V., FREDERICK, P. D. and KIEFE, C. I. (2000). Prevalence, clinical characteristics, and mortality among patients with myocardial infarction presenting without chest pain. *JAMA* **283** 3223–3229.

CARONE, M., ASGHARIAN, M. and WANG, M.-C. (2012). Nonparametric incidence estimation from prevalent cohort survival data. *Biometrika* **99** 599–613. MR2966772 https://doi.org/10.1093/biomet/ass017

Key words and phrases. Functional clustering, disease prevalence, temporal trends, NHIRD.

CHAO, D. Y., CHIEN, Y. Z., YEH, Y. P., HSU, P. S. and LIAN, I. B. (2009). The incidence of varicella and herpes zoster in Taiwan during a period of increasing varicella vaccine coverage, 2000-2008. *Epidemiol. Infect.* **140** 1131–1140.

CHEN, L. L., BLUMM, N., CHRISTAKIS, N. A., BARABASI, A. L. and DEISBOECK, T. S. (2009). Cancer metastasis networks and the prediction of progression patterns. *Br. J. Cancer* **101** 749–758.

CHIANG, C.-J., CHEN, Y.-C., CHEN, C.-J., YOU, S.-L., LAI, M.-S. and TAIWAN CANCER REGISTRY TASK FORCE (2010). Cancer trends in Taiwan. *Jpn. J. Clin. Oncol.* **40** 897–904. https://doi.org/10.1093/jjco/hyq057

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

JACQUES, J. and PREDA, C. (2013). Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing* **112** 164–171.

JACQUES, J. and PREDA, C. (2014). Functional data clustering: A survey. *Adv. Data Anal. Classif.* **8** 231–255. MR3253859 https://doi.org/10.1007/s11634-013-0158-y

JADHAV, S., MA, C., JIANG, Y., SHIA, B.-C. and MA, S. (2021). Supplement to "Pan-disease clustering analysis of the trend of period prevalence." https://doi.org/10.1214/21-AOAS1470SUPPA, https://doi.org/10.1214/21-AOAS1470SUPPB

KAFADAR, K. and KARON, J. M. (1993). An analysis of AIDS incidence data by clustering trends. *Stat. Med.* **12** 311–326.

KEIDING, N. (1991). Age-specific incidence and prevalence: A statistical perspective. *J. Roy. Statist. Soc. Ser. A* **154** 371–412. MR1144166 https://doi.org/10.2307/2983150

KOTHIWALA, S. K., KHANNA, N., TANDON, N., NAIK, N., SHARMA, V. K., SHARMA, S. and SREENIVAS, V. (2016). Prevalence of metabolic syndrome and cardiovascular changes in patients with chronic plaque psoriasis and their correlation with disease severity: A hospital-based cross-sectional study. *Indian J. Dermatol. Venereol. Leprol.* **82** 510.

MA, S. and HUANG, J. (2017). A concave pairwise fusion approach to subgroup analysis. *J. Amer. Statist. Assoc.* **112** 410–423. MR3646581 https://doi.org/10.1080/01621459.2016.1148039

NUTT, D., WILSON, S. and PATERSON, L. (2008). Sleep disorders as core symptoms of depression. *Dialogues Clin. Neurosci.* **10** 329–336.

OHMORI, M., ISHIKAWA, N., YOSHIYAMA, T., UCHIMURA, K., AOKI, M. and MORI, T. (2002). Current epidemiological trend of tuberculosis in Japan. *Int. J. Tuberc. Lung Dis.* **6** 415–423.

PANG, C., GUAN, Y., LI, H., CHEN, W. and ZHU, G. (2016). Urologic cancer in China. *Jpn. J. Clin. Oncol.* **46** 497–501.

PENG, J. and MÜLLER, H.-G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Ann. Appl. Stat.* **2** 1056–1077. MR2516804 https://doi.org/10.1214/08-AOAS172

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2168993

ROTHMAN, K. J. (2012). In Epidemiology: An introduction 37–109, Springer, New York.

RZHETSKY, A., WAJNGURT, D., PARK, N. and ZHENG, T. (2007). Probing genetic overlap among complex human phenotypes. *Proc. Natl. Acad. Sci. USA* **104** 11694–11699.

SEYMOUR, J. M., SPRUIT, M. A., HOPKINSON, N. S., NATANEK, S. A., MAN, W. C., JACKSON, A., GOSKER, H. R., SCHOLS, A. M. W. J., MOXHAM, J. et al. (2010). The prevalence of quadriceps weakness in COPD and the relationship with disease severity. *Eur. Respir. J.* **36** 81–88.

SHEN, X. and HUANG, H.-C. (2010). Grouping pursuit through a regularization solution surface. *J. Amer. Statist. Assoc.* **105** 727–739. MR2724856 https://doi.org/10.1198/jasa.2010.tm09380

TANG, L. and SONG, P. X. K. (2016). Fused lasso approach in regression coefficients clustering—learning parameter heterogeneity in data integration. *J. Mach. Learn. Res.* **17** Paper No. 113, 23. MR3543519

VIEGAS, M., COSTA, C., LOPES, A., GRIZ, L., MEDEIRO, M. A. and BANDEIRA, F. (2011). Prevalence of osteoporosis and vertebral fractures in postmenopausal women with type 2 diabetes mellitus and their relationship with duration of the disease and chronic complications. *J. Diabetes Complicat.* **25** 216–221.

WALTERS, A. S. and RYE, D. B. (2009). Review of the relationship of restless legs syndrome and periodic limb movements in sleep to hypertension, heart disease, and stroke. *Sleep* **32** 589–597.

WEI, W. Q., BASTARACHE, L. A., CARROLL, R. J., MARLO, J. E., OSTERMAN, T. J., GAMAZON, E. R., COX, N. J., RODEN, D. M. and DENNY, J. C. (2007). Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS ONE* **12** e0175508.

YOUNG, A., KODURI, G., BATLEY, M., KULINSKAYA, E., GOUGH, A., NORTON, S. and DIXEY, J. (2006). Mortality in rheumatoid arthritis. Increased in the early course of disease, in ischaemic heart disease and in pulmonary fibrosis. *Rheumatol.* **46** 350–357.

ZHOU, X., LEI, L., LIU, J., HALU, A., ZHANG, Y., LI, B., GUO, Z., LIU, G., SUN, C. et al. (2018). A systems approach to refine disease taxonomy by integrating phenotypic and molecular networks. *EBioMedicine* **31** 79–91.

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1959–1979
https://doi.org/10.1214/21-AOAS1474

SPACE-TIME SMOOTHING MODELS FOR SUBNATIONAL MEASLES ROUTINE IMMUNIZATION COVERAGE ESTIMATION WITH COMPLEX SURVEY DATA

BY TRACY QI DONG[1,*], AND JON WAKEFIELD[2,†]

[1]*Department of Biostatistics, University of Washington,* [*]*qd8@uw.edu*
[2]*Departments of Biostatistics and Statistics, University of Washington,* [†]*jonno@uw.edu*

Despite substantial advances in global measles vaccination, measles disease burden remains high in many low- and middle-income countries. A key public health strategy for controlling measles in such high-burden settings is to conduct supplementary immunization activities (SIAs) in the form of mass vaccination campaigns, in addition to delivering scheduled vaccination through routine immunization (RI) programs. To achieve balanced implementations of RI and SIAs, robust measurement of subnational RI-specific coverage is crucial. In this paper we develop a space–time smoothing model for estimating RI-specific coverage of the first dose of measles-containing-vaccines (MCV1) at subnational level using complex survey data. The application that motivated this work is estimation of the RI-specific MCV1 coverage in Nigeria's 36 states and the Federal Capital Territory. Data come from four demographic and health surveys, three multiple indicator cluster surveys and two national nutrition and health surveys conducted in Nigeria between 2003 and 2018. Our method incorporates information from the SIA calendar published by the World Health Organization and accounts for the impact of SIAs on the overall MCV1 coverage, as measured by cross-sectional surveys. The model can be used to analyze data from multiple surveys with different data collection schemes and construct coverage estimates with uncertainty that reflects the various sampling designs. Implementation of our method can be done efficiently using integrated nested Laplace approximation (INLA).

REFERENCES

BESAG, J., YORK, J. and MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43** 1–20.

BID INITIATIVE (2016). Empowering countries to enhance immunization and overall health service delivery through improved data collection, quality, and use. Available at http://bidinitiative.org/wp-content/uploads/BID_GlobalFactSheet.pdf. [Accessed Apr-02-2020].

BIELLIK, R. J. and ORENSTEIN, W. A. (2018). Strengthening routine immunization through measles-rubella elimination. *Vaccine* **36** 5645–5650.

BINDER, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.* **51** 279–292.

BINYARUKA, P. and BORGHI, J. (2018). Validity of parental recalls to estimate vaccination coverage: Evidence from Tanzania. *BMC Health Serv. Res.* **18** 440.

CUTTS, F. T., CLAQUIN, P., DANOVARO-HOLLIDAY, M. C. and RHODA, D. A. (2016). Monitoring vaccination coverage: Defining the role of surveys. *Vaccine* **34** 4103–4109.

DOLAN, S. B. and MACNEIL, A. (2017). Comparison of inflation of third dose diphtheria tetanus pertussis (DTP3) administrative coverage to other vaccine antigens. *Vaccine* **35** 3441–3445.

DONG, T. Q., RHODA, D. A. and MERCER, L. D. (2020). Impact of state weights on national vaccination coverage estimates from household surveys in Nigeria. *Vaccine* **38** 5060–5070.

DONG, T. Q. and WAKEFIELD, J. (2020). Estimating efficacy of measles supplementary immunization activities via discrete-time modeling of disease incidence time series. Preprint. Available at arXiv:2010.08875.

Key words and phrases. Bayesian smoothing, survey sampling, measles vaccination, routine immunization, supplementary immunization activity.

DONG, T. Q. and WAKEFIELD, J. (2021a). Supplement to "Space–time smoothing models for subnational measles routine immunization coverage estimation with complex survey data." https://doi.org/10.1214/21-AOAS1474SUPP

DONG, T. Q. and WAKEFIELD, J. (2021b). Modeling and presentation of vaccination coverage estimates using data from household surveys. *Vaccine* **39** 2584–2594.

FAY, R. E. III and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277.

HANCIOGLU, A. and ARNOLD, F. (2013). Measuring coverage in MNCH: Tracking progress in health for women and children using DHS and MICS household surveys. *PLoS Med.* **10** e1001391.

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685.

HU, Y., CHEN, Y., WANG, Y. and LIANG, H. (2019). Validity of maternal recall to assess vaccination coverage: Evidence from six districts in Zhejiang province, China. *Int. J. Environ. Res. Public Health* **16** 957.

ICF (2020). The Demographic and Health Surveys. The DHS Program website. Funded by USAID. Available at https://dhsprogram.com/. [Accessed Apr-02-2020].

KNORR-HELD, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Stat. Med.* **19** 2555–2567.

LOCAL BURDEN OF DISEASE VACCINE COVERAGE COLLABORATORS AND OTHERS (2021). Mapping routine measles vaccination in low-and middle-income countries. *Nature* **589** 415–419.

LUMLEY, T. (2004). Analysis of complex survey samples. *J. Stat. Softw.* **9** 1–19.

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

PAIGE, J., FUGLSTAD, G. A., RIEBLER, A. and WAKEFIELD, J. (2020). Model-based approaches to analysing spatial data from complex surveys. *Journal of Survey Statistics and Methodology*. Published online September 4, 2020.

PFEFFERMANN, D., SKINNER, C. J., HOLMES, D. J., GOLDSTEIN, H. and RASBASH, J. (1998). Weighting for unequal selection probabilities in multilevel models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 23–40.

PORTH, J. M., WAGNER, A. L., TEFERA, Y. A. and BOULTON, M. L. (2019). Childhood immunization in Ethiopia: Accuracy of maternal recall compared to vaccination cards. *Vaccines* **7** 48.

R Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

RABE-HESKETH, S. and SKRONDAL, A. (2006). Multilevel modelling of complex survey data. *J. Roy. Statist. Soc. Ser. A* **169** 805–827.

RAMAKRISHNAN, R., VENKATA RAO, T., SUNDARAMOORTHY, L. and JOSHUA, V. (1999). Magnitude of recall bias in the estimation of immunization coverage and its determinants. *Indian Pediatrics* **36** 881–886.

RAO, J., VERRET, F. and HIDIROGLOU, M. A. (2013). A weighted composite likelihood approach to inference for two-level models from survey data. *Surv. Methodol.* **39** 263–282.

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.

SIMPSON, D., RUE, H., RIEBLER, A., MARTINS, T. G. and SØRBYE, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.* **32** 1–28.

SMART TECHNICAL ADVISORY GROUP (2006). Measuring Mortality, Nutritional Status, and Food Security in Crisis Situations: SMART Methodology Manual v1.0. SMART, Action Against Hunger Canada. Available at https://smartmethodology.org/survey-planning-tools/smart-methodology/.

SMART TECHNICAL ADVISORY GROUP (2017). Measuring Mortality, Nutritional Status, and Food Security in Crisis Situations: SMART Methodology Manual v2.0. SMART, Action Against Hunger Canada. Available at https://smartmethodology.org/survey-planning-tools/smart-methodology/.

THAKKAR, N., GILANI, S. S. A., HASAN, Q. and MCCARTHY, K. A. (2019). Decreasing measles burden by optimizing campaign timing. *Proc. Natl. Acad. Sci. USA* **116** 11069–11073.

THE MEASLES AND RUBELLA INITIATIVE (2012). Global Measles and Rubella Strategic Plan 2012–2020. Available at https://measlesrubellainitiative.org/learn/the-solution/the-strategy/. [Accessed Apr-02-2020].

THE MEASLES AND RUBELLA INITIATIVE (2019). Routine Immunization. Available at https://measlesrubellainitiative.org/learn/the-impact/routine-immunization/. [Accessed Apr-02-2020].

UNICEF (2020). The UNICEF Multiple Indicator Cluster Surveys. Available at https://mics.unicef.org/. [Accessed Apr-02-2020].

UTAZI, C. E., THORLEY, J., ALEGANA, V. A., FERRARI, M. J., TAKAHASHI, S., METCALF, C. J. E., LESSLER, J. and TATEM, A. J. (2018). High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. *Vaccine* **36** 1583–1591.

UTAZI, C. E., THORLEY, J., ALEGANA, V. A., FERRARI, M. J., NILSEN, K., TAKAHASHI, S., METCALF, C. J. E., LESSLER, J. and TATEM, A. J. (2019a). A spatial regression model for the disaggregation of areal unit based data to high-resolution grids with application to vaccination coverage mapping. *Stat. Methods Med. Res.* **28** 3226–3241.

UTAZI, C. E., THORLEY, J., ALEGANA, V. A., FERRARI, M. J., TAKAHASHI, S., METCALF, C. J. E., LESSLER, J., CUTTS, F. T. and TATEM, A. J. (2019b). Mapping vaccination coverage to explore the effects of delivery mechanisms and inform vaccination strategies. *Nat. Commun.* **10** 1633.

UTAZI, C. E., WAGAI, J., PANNELL, O., CUTTS, F. T., RHODA, D. A., FERRARI, M. J., DIENG, B., OTERI, J., DANOVARO-HOLLIDAY, M. C. et al. (2020). Geospatial variation in measles vaccine coverage through routine and campaign strategies in Nigeria: Analysis of recent household surveys. *Vaccine*. **38** 3062–3071.

VALADEZ, J. J. and WELD, L. H. (1992). Maternal recall error of child vaccination status in a developing nation. *Am. J. Publ. Health* **82** 120–122.

VERGUET, S., JOHRI, M., MORRIS, S. K., GAUVREAU, C. L., JHA, P. and JIT, M. (2015). Controlling measles using supplemental immunization activities: A mathematical model to inform optimal policy. *Vaccine* **33** 1291–1296.

WILSON, K. and WAKEFIELD, J. (2020). Child mortality estimation incorporating summary birth history data. *Biometrics* 1–11.

WORLD HEALTH ORGANIZATION (2017). Measles position paper. Available at https://www.who.int/immunization/policy/position_papers/measles/en/. [Accessed Apr-02-2020].

WORLD HEALTH ORGANIZATION (2020a). Immunization Agenda 2030: A Global Strategy to Leave No One Behind. Available at https://www.who.int/immunization/immunization_agenda_2030/en/. [Accessed Apr-07-2020].

WORLD HEALTH ORGANIZATION (2020b). Immunization, vaccines and biologicals—Data, statistics and graphics. Available at https://www.who.int/immunization/monitoring_surveillance/data/en/. [Accessed Apr-07-2020].

WORLD HEALTH ORGANIZATION (2020c). At least 80 million children under one at risk of diseases such as diphtheria, measles and polio as COVID-19 disrupts routine vaccination efforts, warn Gavi, WHO and UNICEF. Available at https://www.who.int/news-room/detail/22-05-2020-at-least-80-million-children-under-one-at-risk-of-diseases-such-as-diphtheria-measles-and-polio-as-covid-19-disrupts-routine-vaccination-efforts-warn-gavi-who-and-unicef. [Accessed Jun-28-2020].

YI, G. Y., RAO, J. N. K. and LI, H. (2016). A weighted composite likelihood approach for analysis of survey data under two-level models. *Statist. Sinica* **26** 569–587.

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1980–1998
https://doi.org/10.1214/21-AOAS1486

INTEGRATING GEOSTATISTICAL MAPS AND INFECTIOUS DISEASE TRANSMISSION MODELS USING ADAPTIVE MULTIPLE IMPORTANCE SAMPLING

BY RENATA RETKUTE[1], PANAYIOTA TOULOUPOU[2], MARÍA-GLORIA BASÁÑEZ[3], T. DÉIRDRE HOLLINGSWORTH[4] AND SIMON E. F. SPENCER[5]

[1]*Epidemiology and Modelling Group, Department of Plant Sciences, University of Cambridge, rr614@cam.ac.uk*
[2]*School of Mathematics, University of Birmingham, p.touloupou@bham.ac.uk*
[3]*London Centre for Neglected Tropical Disease Research and MRC Centre for Global Infectious Disease Analysis, Faculty of Medicine, School of Public Health, Imperial College London, m.basanez@imperial.ac.uk*
[4]*Big Data Institute, Li Ka Shing Centre for Health, Information and Discovery, University of Oxford, deirdre.hollingsworth@bdi.ox.ac.uk*
[5]*Department of Statistics and Zeeman Institute, University of Warwick, s.e.f.spencer@warwick.ac.uk*

The Adaptive Multiple Importance Sampling algorithm (AMIS) is an iterative technique which recycles samples from all previous iterations in order to improve the efficiency of the proposal distribution. We have formulated a new statistical framework, based on AMIS, to take the output from a geostatistical model of infectious disease prevalence, incidence or relative risk, and project it forward in time under a mathematical model for transmission dynamics. We adapted the AMIS algorithm so that it can sample from multiple targets simultaneously by changing the focus of the adaptation at each iteration. By comparing our approach against the standard AMIS algorithm, we showed that these novel adaptations greatly improve the efficiency of the sampling. We tested the performance of our algorithm on four case studies: ascariasis in Ethiopia, onchocerciasis in Togo, human immunodeficiency virus (HIV) in Botswana, and malaria in the Democratic Republic of the Congo.

REFERENCES

ALKEMA, L., RAFTERY, A. E. and CLARK, S. J. (2007). Probabilistic projections of HIV prevalence using Bayesian melding. *Ann. Appl. Stat.* **1** 229–248. MR2393849 https://doi.org/10.1214/07-AOAS111

AMOAH, B., DIGGLE, P. J. and GIORGI, E. (2020). A geostatistical framework for combining spatially referenced disease prevalence data from multiple diagnostics. *Biometrics* **76** 158–170. MR4098552 https://doi.org/10.1111/biom.13142

ANDERSON, R. M. and BASÁÑEZ, M. G., eds. (2015). Adv Parasitol., Part A. In *Mathematical Models for Neglected Tropical Diseases*: *Essential Tools for Control and Elimination* **87** Academic Press, Oxford.

ANDERSON, R. M. and MAY, R. M. (1985). Helminth infections of humans: Mathematical models, population dynamics, and control. *Adv. Parasitol.* **24** 1–101. https://doi.org/10.1016/s0065-308x(08)60561-8

ANDERSON, R. A. and MAY, R. M. (1992). *Infectious Diseases of Humans*, 1st ed. ed. Oxford Univ. Press, Oxford.

BASÁÑEZ, M. G. and ANDERSON, R. M., eds. (2016). Adv Parasitol., Part B. In *Mathematical Models for Neglected Tropical Diseases*: *Essential Tools for Control and Elimination* **94** Academic Press, Oxford.

BHATT, S., GETHING, P. W., BRADY, O. J., MESSINA, J. P., FARLOW, A. W., MOYES, C. L., DRAKE, J. M., BROWNSTEIN, J. S., HOEN, A. G. et al. (2013). The global distribution and burden of Dengue. *Nature* **496** 504–507.

BHATT, S., WEISS, D. J., CAMERON, E., BISANZIO, D., MAPPIN, B., DALRYMPLE, U., BATTLE, K., MOYES, C. L., HENRY, A. et al. (2015). The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature* **526** 207–211. https://doi.org/10.1038/nature15535

BROOKER, S., CLEMENTS, A. C. A. and BUNDY, D. A. P. (2006). Global epidemiology, ecology and control of soil-transmitted helminth infections. *Adv. Parasitol.* **62** 221–261.

CAPPÉ, O., DOUC, R., GUILLIN, A., MARIN, J.-M. and ROBERT, C. P. (2008). Adaptive importance sampling in general mixture classes. *Stat. Comput.* **18** 447–459. MR2461888 https://doi.org/10.1007/s11222-008-9059-x

Key words and phrases. Epidemiology, Disease mapping, Parameter estimation, Importance sampling.

CHEKE, R. A., BASÁÑEZ, M.-G., PERRY, M., WHITE, M. T., GARMS, R., OBUOBIE, E., LAMBERTON, P. H. L., YOUNG, S., OSEI-ATWENEBOANA, M. Y. et al. (2015). Potential effects of warmer worms and vectors on onchocerciasis transmission in West Africa. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **370**. https://doi.org/10.1098/rstb.2013.0559

COLEBUNDERS, R., BASÁÑEZ, M.-G., SILING, K., POST, R. J., ROTSAERT, A., MMBANDO, B., SUYKERBUYK, P. and HOPKINS, A. (2018). From river blindness control to elimination: Bridge over troubled water. *Infect. Dis. Poverty* **7** 21. https://doi.org/10.1186/s40249-018-0406-7

CORNUET, J.-M., MARIN, J.-M., MIRA, A. and ROBERT, C. P. (2012). Adaptive multiple importance sampling. *Scand. J. Stat.* **39** 798–812. MR3000850 https://doi.org/10.1111/j.1467-9469.2011.00756.x

DIGGLE, P. and GIORGI, E. (2019). *Model-Based Geostatistics for Global Public Health*: *Methods and Applications*. CRC Press, London.

DWYER-LINDGREN, L., CORK, M. A., SLIGAR, A., STEUBEN, K. M., WILSON, K. F., PROVOST, N. R., MAYALA, B. K., VANDER-HEIDE, J. D., COLLISON, M. L. et al. (2019). Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017. *Nature* **570** 189–193.

ELVIRA, V. and SANTAMARIA, I. (2019). Multiple importance sampling for efficient symbol error rate estimation. *IEEE Signal Process. Lett.* **2** 420–424.

FOWLER, A. C. and HOLLINGSWORTH, T. D. (2016). The dynamics of *Ascaris lumbricoides* infections. *Bull. Math. Biol.* **78** 815–833. MR3494574 https://doi.org/10.1007/s11538-016-0164-2

FRALEY, C. and RAFTERY, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J.* **41** 578–588.

GIORGI, E., DIGGLE, P. J., SNOW, R. W. and NOOR, A. M. (2018). Geostatistical methods for disease mapping and visualisation using data from spatio-temporally referenced prevalence surveys. *Int. Stat. Rev.* **86** 571–597. MR3882131 https://doi.org/10.1111/insr.12268

GUYATT, H. L., BUNDY, D. A., MEDLEY, G. F. and GRENFELL, B. T. (1990). The relationship between the frequency distribution of *Ascaris lumbricoides* and the prevalence and intensity of infection in human communities. *Parasitology* **101** 139–143.

HAMLEY, J. I. D., MILTON, P., WALKER, M. and BASÁÑEZ, M. G. (2019). Modelling exposure heterogeneity and density dependence in onchocerciasis using a novel individual-based transmission model, EPIONCHO-IBM: Implications for elimination and data needs. *PLoS Negl. Trop. Dis.*. **13** e0007557.

HAY, S. I., BATTLE, K. E., PIGOTT, D. M., SMITH, D. L., MOYES, C. L., BHATT, S., BROWNSTEIN, J. S., COLLIER, N., MYERS, M. F. et al. (2013). Global mapping of infectious disease. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **368** 20120250. https://doi.org/10.1098/rstb.2012.0250

HIJMANS, R. J. (2019). Raster: Geographic data analysis and modeling. R package version 2.4-15. http://CRAN.R-project.org/package=raster.

HOLLINGSWORTH, T. D. (2018). Counting down the 2020 goals for 9 neglected tropical diseases: What have we learned from quantitative analysis and transmission modeling? *Clin. Infect. Dis.* **66** S237–S244.

HOUGARD, J. M., ALLEY, E. S., YAMÉOGO, L., DADZIE, K. Y. and BOATIN, B. A. (2001). Eliminating onchocerciasis after 14 years of vector control: A proved strategy. *J. Infect. Dis.* **184** 497–503. https://doi.org/10.1086/322789

JACOB, B. G., LOUM, D., LAKWO, T. L., KATHOLI, C. R., HABOMUGISHA, P., BYAMUKAMA, E., TUKAHEBWA, E., CUPP, E. W. and UNNASCH, T. R. (2018). Community-directed vector control to supplement mass drug distribution for onchocerciasis elimination in the Madi mid-North focus of northern Uganda. *PLoS Negl. Trop. Dis.* **12** e0006702. https://doi.org/10.1371/journal.pntd.0006702

KARAGIANNIS-VOULES, D.-A., BIEDERMANN, P., EKPO, U. F., GARBA, A., LANGER, E., MATHIEU, E., MIDZI, N., MWINZI, P., POLDERMAN, A. M. et al. (2015). Spatial and temporal distribution of soil-transmitted helminth infection in sub-Saharan Africa: A systematic review and geostatistical meta-analysis. *Lancet Infect. Dis.* **15** 74–84. https://doi.org/10.1016/S1473-3099(14)71004-7

KEELING, M. J. and ROHANI, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton Univ. Press, Princeton, NJ. MR2354763

KISH, L. (1965). *Survey Sampling*, 1st ed. ed. Wiley, New York.

LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. MR2853727 https://doi.org/10.1111/j.1467-9868.2011.00777.x

MARIN, J.-M., PUDLO, P. and SEDKI, M. (2019). Consistency of adaptive importance sampling and recycling schemes. *Bernoulli* **25** 1977–1998. MR3961237 https://doi.org/10.3150/18-BEJ1042

MARTINO, L., ELVIRA, V., LUENGO, D. and CORANDER, J. (2017). Layered adaptive importance sampling. *Stat. Comput.* **27** 599–623. MR3613588 https://doi.org/10.1007/s11222-016-9642-5

MICHAEL, E., SMITH, M. E., KATABARWA, M. N., BYAMUKAMA, E., GRISWOLD, E., HABOMUGISHA, P., LAKWO, T., TUKAHEBWA, E., MIRI, E. S. et al. (2018). Substantiating freedom from parasitic infection by combining transmission model predictions with disease surveys. *Nat. Commun.* **9** 4324.

MOSSER, J. F., GAGNE-MAYNARD, W., RAO, P. C., OSGOOD-ZIMMERMAN, A., FULLMAN, N., GRAETZ, N., BURSTEIN, R., UPDIKE, R. L., LIU, P. Y. et al. (2019). Mapping diphtheria-pertussis-tetanus vaccine coverage in Africa, 2000–2016: A spatial and temporal modelling study. *Lancet* **4** 1843–1855.

NORHAYATI, M., FATMAH, M. S., YUSOF, S. and EDARIAH, A. B. (2003). Intestinal parasitic infections in man: A review. *Med. J. Malays.* **58** 296–305.

NTAMABYALIRO, N. Y., BURRI, C., NZOLO, D. B., ENGO, A. B., LULA, Y. N., MAMPUNZA, S. M., NSIBU, C. N., MESIA, G. K., KAYEMBE, J. M. N. et al. (2018). Drug use in the management of uncomplicated malaria in public health facilities in the Democratic Republic of the Congo. *Malar. J.* **17** 189.

O'HANLON, S. J., SLATER, H. C., CHEKE, R. A., BOATIN, B. A., COFFENG, L. E., PION, S. D. S., BOUSSINESQ, M., ZOURE, H. G. M., STOLK, W. A. et al. (2016). Model-based geostatistical mapping of the prevalence of *Onchocerca volvulus* in West Africa. *PLoS Negl. Trop. Dis.* **10** e0004328.

PIGOTT, D. M., BHATT, S., GOLDING, N., DUDA, K. A., BATTLE, K. E., BRADY, O. J., MESSINA, J. P., BALARD, Y., BASTIEN, P. et al. (2014). Global distribution maps of the leishmaniases. *eLife* **3** e02851.

PLAISIER, A. P., VAN OORTMARSSEN, G. J., REMME, J. and HABBEMA, J. P. (1991). The reproductive lifespan of *Onchocerca volvulus* in West African savanna. *Acta Trop.* **48** 271–284.

POOLE, D. and RAFTERY, A. E. (2000). Inference for deterministic simulation models: The Bayesian melding approach. *J. Amer. Statist. Assoc.* **95** 1244–1255. MR1804247 https://doi.org/10.2307/2669764

PULLAN, R. L., SMITH, J. L., JASRASARIA, R. and BROOKER, S. J. (2014). Global numbers of infection and disease burden of soil transmitted helminth infections in 2010. *Parasit. Vectors BioMed. Central.* **7** 37.

RAFTERY, A. E. and BAO, L. (2010). Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. *Biometrics* **66** 1162–1173. MR2758504 https://doi.org/10.1111/j.1541-0420.2010.01399.x

RETKUTE, R., TOULOUPOU, P., BASÁÑEZ, M.-G., HOLLINGSWORTH, T. D. and SPENCER, S. E. (2021). Supplement to "Integrating geostatistical maps and infectious disease transmission models using adaptive multiple importance sampling." https://doi.org/10.1214/21-AOAS1486SUPPA, https://doi.org/10.1214/21-AOAS1486SUPPB

RIPLEY, B. D. (1987). *Stochastic Simulation*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0875224 https://doi.org/10.1002/9780470316726

ROUTLEDGE, I., WALKER, M., CHEKE, R. A., BHATT, S., NKOT, P. B., MATTHEWS, G. A., BALEGUEL, D., DOBSON, H. M., WILES, T. L. et al. (2018). Modelling the impact of larviciding on the population dynamics and biting rates of *Simulium damnosum* (s.l.): Implications for vector control as a complementary strategy for onchocerciasis elimination in Africa. *Parasites Vectors* **11** 316. https://doi.org/10.1186/s13071-018-2864-y

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

SBERT, M. and HAVRAN, V. (2017). Adaptive multiple importance sampling for general functions. *Vis. Comput.* **33** 6–8.

SCRUCCA, L., FOP, M., MURPHY, T. B. and RAFTERY, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8** 289–317.

SIMARRO, P. P., CECCHI, G., FRANCO, J. R., PAONE, M., DIARRA, A., RUIZ-POSTIGO, J. A., FÈVRE, E. M., MATTIOLI, R. C. and JANNIN, J. G. (2012). Estimating and mapping the population at risk of sleeping sickness. *PLoS Negl. Trop. Dis.* **6** e1859. https://doi.org/10.1371/journal.pntd.0001859

SIREN, J., MARTTINEN, P. and CORANDER, J. (2010). Reconstructing population histories from single-nucleotide polymorphism data. *Mol. Biol. Evol.* **28** 673–683.

TOULOUPOU, P., RETKUTE, R., HOLLINGSWORTH, T. D. and SPENCER, S. E. F. (2020). Statistical methods for linking geostatistical maps and transmission models: Application to lymphatic filariasis in East Africa. *Spat. Spatio-Tempor. Epidemiol.* **100391** 1–10.

VEACH, E. and GUIBAS, L. (1995). Optimally combining sampling techniques for Monte Carlo rendering. In *Proceedings of the* 22*nd Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH* **95** 419–428.

VERVER, S., WALKER, M., KIM, Y. E., FOBI, G., TEKLE, A. H., ZOURE, H. G. M., WANJI, S., BOAKYE, D. A., KUESEL, A. C. et al. (2018). How can onchocerciasis elimination in Africa be accelerated? Modeling the impact of increased ivermectin treatment frequency and complementary vector control. *Clin. Infect. Dis.*. **66** S267–S274.

WALKER, M., STOLK, W. A., DIXON, M. A., BOTTOMLEY, C., DIAWARA, L., TRAORÉ, M. O., DE VLAS, S. J. and BASÁÑEZ, M. G. (2017). Modelling the elimination of river blindness using long-term epidemiological and programmatic data from Mali and Senegal. *Epidemics* **18** 4–15. https://doi.org/10.1016/j.epidem.2017.02.005

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1999–2022
https://doi.org/10.1214/21-AOAS1450

ANALYSING THE CAUSAL EFFECT OF LONDON CYCLE SUPERHIGHWAYS ON TRAFFIC CONGESTION

BY PRAJAMITRA BHUYAN[1,*], EMMA J. MCCOY[1,†], HAOJIE LI[2] AND DANIEL J. GRAHAM[3]

[1]*Department of Mathematics, Imperial College London,* [*]*p.bhuyan@imperial.ac.uk;* [†]*e.mccoy@imperial.ac.uk*
[2]*School of Transportation, Southeast University, h.li@seu.edu.cn*
[3]*Department of Civil and Environmental Engineering, Imperial College London, d.j.graham@imperial.ac.uk*

Transport operators have a range of intervention options available to improve or enhance their networks. Such interventions are often made in the absence of sound evidence on resulting outcomes. Cycling superhighways were promoted as a sustainable and healthy travel mode, one of the aims of which was to reduce traffic congestion. Estimating the impacts that cycle superhighways have on congestion is complicated due to the nonrandom assignment of such intervention over the transport network. In this paper we analyse the causal effect of cycle superhighways utilising preintervention and postintervention information on traffic and road characteristics along with socioeconomic factors. We propose a modeling framework based on the propensity score and outcome regression model. The method is also extended to the doubly robust set-up. Simulation results show the superiority of the performance of the proposed method over existing competitors. The method is applied to analyse a real dataset on the London transport network. The methodology proposed can assist in effective decision making to improve network performance.

REFERENCES

ABADIE, A. (2005). Semiparametric difference-in-differences estimators. *Rev. Econ. Stud.* **72** 1–19. MR2116973 https://doi.org/10.1111/0034-6527.00321

ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267. MR2194325 https://doi.org/10.1111/j.1468-0262.2006.00655.x

AUSTIN, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* **46** 399–424. https://doi.org/10.1080/00273171.2011.568786

BADOE, D. A. and MILLER, E. J. (2000). Transportation-land-use interaction: Empirical findings in North America, and their implications for modeling. *Transp. Res., Part D, Transp. Environ.* **5** 235–263.

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189 https://doi.org/10.1111/j.1541-0420.2005.00377.x

BARKLEY, B. G., HUDGENS, M. G., CLEMENS, J. D., ALI, M. and EMCH, M. E. (2020). Causal inference from observational studies with clustered interference, with application to a cholera vaccine study. *Ann. Appl. Stat.* **14** 1432–1448. MR4152140 https://doi.org/10.1214/19-AOAS1314

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). High-dimensional methods and inference on structural and treatment effects. *J. Econ. Perspect.* **28** 29–50.

BELLONI, A., CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and HANSEN, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85** 233–298. MR3611771 https://doi.org/10.3982/ECTA12723

BLUNDEN, M. (2016). Cycle superhighways make traffic worse in the city, report reveals. *Evening Stand.* **Oct 5**.

BRANAS, C. C., CHENEY, R. A., MACDONALD, J. M., TAM, V. W., JACKSON, T. D. and TEN HAVE, T. R. (2011). A difference-in-differences analysis of health, safety, and greening vacant urban space. *Am. J. Epidemiol.* **174** 1296–1306.

CALLAWAY, B. and SANT'ANNA, P. H. C. (2020). Difference-in-differences with multiple time periods. *J. Econometrics*.

Key words and phrases. Average treatment effect, confounder, difference-in-difference, intelligent transportation system, potential outcome.

CARD, D. and KRUEGER, A. B. (1994). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. *Am. Econ. Rev.* **84** 772–93.
CHAISEMARTIN, D. and D'HAULTFOEUILLE, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *Am. Econ. Rev.* Available at https://arxiv.org/abs/1803.08807.
CHEN, C., JIA, Z. and VARAIYA, P. (2001). Causes and cures of highway congestion. *IEEE Control Syst. Mag.* **21** 26–32.
DANIEL, R. M., COUSENS, S. N., DE STAVOLA, B. L., KENWARD, M. G. and STERNE, J. A. C. (2013). Methods for dealing with time-dependent confounding. *Stat. Med.* **32** 1584–1618. MR3060620 https://doi.org/10.1002/sim.5686
DAW, J. R. and HATFIELD, L. A. (2018). Matching and regression to the mean in difference-in-differences analysis. *Health Serv. Res.* **53** 4138–4156. https://doi.org/10.1111/1475-6773.12993
DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR2049007
FLORES, C. A., FLORES-LAGUNES, A., GONZALEZ, A. and NEUMANN, T. C. (2012). Estimating the effects of length of exposure to instruction in a training program: The case of job corps. *Rev. Econ. Stat.* **94** 153–171.
GEORGE, K. A. (1970). Transportation compatible land uses and bus-stop location. *Trans. Built Environ.* **44**.
GRAHAM, D. J., MCCOY, E. J. and STEPHENS, D. A. (2016). Approximate Bayesian inference for doubly robust estimation. *Bayesian Anal.* **11** 47–69. MR3447091 https://doi.org/10.1214/14-BA928
HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331. MR1612242 https://doi.org/10.2307/2998560
HARMAN, J. S., LEMAK, C. H., AL-AMIN, M., HALL, A. G. and DUNCAN, R. P. (2011). Changes in per member per month expenditures after implementation of Florida's medicaid reform demonstration. *Health Serv. Res.* **43** 787–804.
HECKMAN, J., ICHIMURA, H., SMITH, J. and TODD, P. (1998). Characterizing selection bias using experimental data. *Econometrica* **66** 1017–1098. MR1639419 https://doi.org/10.2307/2999630
HERNÁN, M. A. and ROBINS, J. M. (2020). *Causal Inference*: *What If*. CRC Press/CRC, Boca Raton.
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 https://doi.org/10.1111/1468-0262.00442
HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460
HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66), *Vol. I*: *Statistics* 221–233. Univ. California Press, Berkeley, Calif. MR0216620
HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292
JIN, J. and RAFFERTY, P. (2017). Does congestion negatively affect income growth and employment growth? Empirical evidence from US metropolitan regions. *Transp. Policy* **55** 1–8.
KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458 https://doi.org/10.1214/07-STS227
LECHNER, L. (2001). The estimation of causal effects by difference-in-difference methods. *Found Trends Econom.* **4** 165–224.
LEE, B. K., LESSLER, J. and STUART, E. A. (2010). Improving propensity score weighting using machine learning. *Stat. Med.* **29** 337–346. MR2750549 https://doi.org/10.1002/sim.3782
LI, H., GRAHAM, D. J. and LIU, P. (2017). Safety effects of the London cycle superhighways on cycle collisions. *Accident Anal. Prev.* **90** 90–101.
LINDNER, S. and MCCONNELL, K. J. (2018). Difference-in-differences and matching on outcomes: A tale of two unobservables. *Health Serv. Outcomes Res. Methodol.* **19** 127–144.
LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **23** 2937–2960. https://doi.org/10.1002/sim.1903
MCCULLOCH, C. E. and NEUHAUS, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statist. Sci.* **26** 388–402. MR2917962 https://doi.org/10.1214/11-STS361
MOODIE, E. E. M., SAARELA, O. and STEPHENS, D. A. (2018). A doubly robust weighting estimator of the average treatment effect on the treated. *Stat* **7** e205. MR3905861 https://doi.org/10.1002/sta4.205
NORMAN, W. (2017). Bike lanes don't clog up our roads, they keep London moving. *The Gaurdian* **Dec 1**.
RETALLACK, A. E. and OSTENDORF, B. (2019). Current understanding of the effects of congestion on traffic accidents. *Int. J. Environ. Res. Public Health* **16**. https://doi.org/10.3390/ijerph16183400
ROBINS, J. M., HERNÁN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

ROBINS, J. M. and ROTNITZKY, A. (2001). A comment on "Inference for semiparametric models: Some questions and an answer". *Statist. Sinica* **11** 920–936.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

SANT'ANNA, P. H. C. and ZHAO, J. (2020). Doubly robust difference-in-differences estimators. *J. Econometrics* **219** 101–122. MR4152787 https://doi.org/10.1016/j.jeconom.2020.06.003

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. With comments and a rejoinder by the authors. MR1731478 https://doi.org/10.2307/2669923

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics. Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0595165

SKRONDAL, A. and RABE-HESKETH, S. (2009). Prediction in multilevel generalized linear models. *J. Roy. Statist. Soc. Ser. A* **172** 659–687. MR2751671 https://doi.org/10.1111/j.1467-985X.2009.00587.x

SLAWSON, N. (2017). Traffic jams on major UK roads cost economy around £9bn. *The Gaurdian* **Oct 18**.

TAN, Z. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Ann. Statist.* **48** 811–837. MR4102677 https://doi.org/10.1214/19-AOS1824

TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21** 55–75. MR2867538 https://doi.org/10.1177/0962280210386779

THE DEPARTMENT OF TRANSPORT (2018). Traffic statistics-methodology review-alternative data sources.

TRANSPORT FOR LONDON (2010). Cycling revolution London.

TRANSPORT FOR LONDON (2011). Barclays cycle superhighways evaluation of pilot routes 3 and 7.

TRANSPORT FOR LONDON (2014). Number of daily cycle journeys in London.

WESTREICH, D., LESSLER, J. and FUNK, M. J. (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J. Clin. Epidemiol.* **63** 826–833.

WHARAM, J. F., LANDON, B. E., GALBRAITH, A. A., KLEINMAN, K. P., SOUMERAI, S. B. and ROSS-DEGNAN, D. (2007). Emergency department use and subsequent hospitalizations among members of a high-deductible health plan. *JAMA* **297** 1093–1102. https://doi.org/10.1001/jama.297.10.1093

YUAN, K., KNOOP, V. L. and HOOGENDOORN, S. P. (2015). Capacity drop: Relationship between speed in congestion and the queue discharge rate. *Transp. Res. Rec.* **2491** 72–80.

ZHANG, K., SUN, D. J., SHEN, S. and ZHU, Y. (2017). Analyzing spatiotemporal congestion pattern on urban roads based on taxi GPS data. *J. Transp. Land Use* **10** 675–694.

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 2023–2042
https://doi.org/10.1214/21-AOAS1448

THE INFORMATION IN COVARIATE IMBALANCE IN STUDIES OF HORMONE REPLACEMENT THERAPY

BY RUOQI YU*, DYLAN S. SMALL† AND PAUL R. ROSENBAUM‡

*Department of Statistics, Wharton School, University of Pennsylvania, *ruoqiyu@wharton.upenn.edu; †dsmall@wharton.upenn.edu; ‡rosenbaum@wharton.upenn.edu*

A widely noted failure of causal inference occurred when several observational studies claimed that hormone replacement therapy (HRT) reduced risk of cardiovascular disease; yet, subsequent randomized trials found an increased, not a decreased, cardiovascular risk. We take a close look at covariate imbalances in one of the observational data sets. We use some old, some recent, and some new methods, plus we update an important, simple but largely forgotten suggestion of William Cochran about screening covariates and other variables. In particular, a tapered match shows the impact on all covariates of gradually matching for additional covariates. An exterior match examines the change in the control group as additional covariates are included, and the consequences for outcomes. Because covariates are sometimes continuous, sometimes binary, sometimes ordinal, sometimes missing, we suggest keeping track of magnitudes of aggregate bias in observed covariates using a new estimate of the Kullback–Leibler information between covariate distributions in treated and matched control groups, a flexible measure with several attractive properties. The initial studies ignored some enormous imbalances in socioeconomic covariates that predict the outcomes under study. Our more comprehensive analyses mimic some post-game reanalyses done subsequent to the randomized trials; however, even these omit a large imbalance in a consequential covariate discovered by Cochran's quick but expansive screening suggestion. Our sense is that a closer examination of covariate imbalance would not have led to a correct conclusion about the effects of HRT, but it would have heightened concerns about the magnitude of the problems in the observational studies, and it would have raised doubts about the ability of a few regression coefficients to eliminate all biases, observed and unobserved, in the comparison. Medical journals need to recognize that certain sources of uncertainty cannot be eliminated from certain necessary types of empirical investigation; moreover, these journals need to learn new ways to describe these sources of uncertainty with objectivity and candor.

REFERENCES

COCHRAN, W. G. (1965). The planning of observational studies of human populations (with discussion). *J. R. Stat. Soc., A* **128** 234–266.

COCHRAN, W. G. and RUBIN, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya, Ser. A* **35** 417–446.

DANIEL, S. R., ARMSTRONG, K., SILBER, J. H. and ROSENBAUM, P. R. (2008). An algorithm for optimal tapered matching, with application to disparities in survival. *J. Comput. Graph. Statist.* **17** 914–924. MR2649074 https://doi.org/10.1198/106186008X385806

JOHANNES, C. B., CRAWFORD, S. L., POSNER, J. G. and MCKINLAY, S. M. (1994). Longitudinal patterns and correlates of hormone replacement therapy use in middle-aged women. *Am. J. Epidemiol.* **140** 439–452.

KELZ, R. R., SELLERS, M. M., NIKNAM, B. A., SHARPE, J. E., ROSENBAUM, P. R., HILL, A. S., ZHOU, H., HOCHMAN, L. L., BILIMORIA, K. Y. et al. (2021). A national comparison of operative outcomes of new and experienced surgeons. *Ann. Surg.* **273** 280–288.

KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, New York. MR0103557

Key words and phrases. Kullback–Leibler information, observational studies, randomized trials, treatment effects.

KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* **22** 79–86. MR0039968 https://doi.org/10.1214/aoms/1177729694

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. MR2135927

MACPHERSON, H., PIPINGAS, A. and PASE, M. P. (2013). Multivitamin–multimineral supplementation and mortality: A meta-analysis of randomized controlled trials. *Am. J. Clin. Nutr.* **97** 437–444.

MATTHEWS, K. A., KULLER, L. H., WING, R. R., MEILAHN, E. N. and PLANTINGA, P. (1996). Prior to use of estrogen replacement therapy, are users healthier than nonusers? *Am. J. Epidemiol.* **143** 971–978.

NIKNAM, B. A., ARRIAGA, A. F., ROSENBAUM, P. R., HILL, A. S., ROSS, R. N., EVEN-SHOSHAN, O., ROMANO, P. S. and SILBER, J. H. (2018). Adjustment for atherosclerosis diagnosis distorts the effects of percutaneous coronary intervention and the ranking of hospital performance. *J. Amer. Heart Assoc.* **7**. https://doi.org/10.1161/JAHA.117.008366

O'BRIEN, P. C. and FLEMING, T. R. (1987). A paired Prentice–Wilcoxon test for censored paired data. *Biometrics* **43** 169–180.

PETITTI, D. B. and FREEDMAN, D. A. (2005). Invited commentary: How far can epidemiologists get with statistical adjustment? *Am. J. Epidemiol.* **162** 415–418.

PETITTI, D. B., PERLMAN, J. A. and SIDNEY, S. (1986). Letter about 'Postmenopausal estrogen use and heart disease'. *N. Engl. J. Med.* **315** 131–132.

PIMENTEL, S. D., SMALL, D. S. and ROSENBAUM, P. R. (2016). Constructed second control groups and attenuation of unmeasured biases. *J. Amer. Statist. Assoc.* **111** 1157–1167. MR3561939 https://doi.org/10.1080/01621459.2015.1076342

PRENTICE, R. L., LANGER, R., STEFANICK, M. L., HOWARD, B. V., PETTINGER, M., ANDERSON, G., BARAD, D., CURB, J. D., KOTCHEN, J. et al. (2005). Combined postmenopausal hormone therapy and cardiovascular disease: Toward resolving the discrepancy between non-experimental studies and the Women's Health Initiative clinical trial. *Am. J. Epidemiol.* **162** 404–420. https://doi.org/10.1093/aje/kwi223

ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. MR0885915 https://doi.org/10.1093/biomet/74.1.13

ROSENBAUM, P. R. (1989). The role of known effects in observational studies. *Biometrics* **45** 557–569. MR1010518 https://doi.org/10.2307/2531497

ROSENBAUM, P. R. (1991). Discussing hidden bias in observational studies. *Ann. Intern. Med.* **115** 901–905.

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1899138 https://doi.org/10.1007/978-1-4757-3692-2

ROSENBAUM, P. R. (2017). *Observation and Experiment*: *An Introduction to Causal Inference*. Harvard Univ. Press, Cambridge, MA. MR3702029

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

ROSENBAUM, P. R. and RUBIN, D. B. (1985). The bias due to incomplete matching. *Biometrics* **41** 103–116. MR0793436 https://doi.org/10.2307/2530647

ROSENBAUM, P. R. and SILBER, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *J. Amer. Statist. Assoc.* **104** 1398–1405. MR2750570 https://doi.org/10.1198/jasa.2009.tm08470

ROSENBAUM, P. R. and SILBER, J. H. (2013). Using the exterior match to compare two entwined matched control groups. *Amer. Statist.* **67** 67–75. MR3303593 https://doi.org/10.1080/00031305.2013.769914

RUTTER, M., ed. (2007). *Identifying the Environmental Causes of Disease*. Academy of Medical Sciences, London.

SILBER, J. H., ROSENBAUM, P. R., CLARK, A. S., GIANTONIO, B. J., ROSS, R. N., TENG, Y., WANG, M., NIKNAM, B. A., LUDWIG, J. M. et al. (2013). Characteristics associated with differences in survival among black and white women with breast cancer. *J. Am. Med. Assoc.* **310** 389–397.

SILBER, J. H., ROSENBAUM, P. R., ROSS, R. N., NIKNAM, B. A., LUDWIG, J. M., WANG, W., CLARK, A. S., FOX, K. R., WANG, M. et al. (2014). Racial disparities in colon cancer survival: A matched cohort study. *Ann. Intern. Med.* **161** 845–854.

SILBER, J. H., ROSENBAUM, P. R., ROSS, R. N., REITER, J. G., NIKNAM, B. A., HILL, A. S., BONGIORNO, D. M., SHAH, S. A., HOCHMAN, L. L. et al. (2018). Disparities in breast cancer survival by socioeconomic status despite medicare and medicaid insurance. *Milbank Q.* **96** 706–754. https://doi.org/10.1111/1468-0009.12355

STAMPFER, M. J., WILLETT, W. C., COLDITZ, G. A., ROSNER, B., SPEIZER, F. E. and HENNEKENS, C. H. (1985). A prospective study of postmenopausal estrogen therapy and coronary heart disease. *N. Engl. J. Med.* **313** 1044–1049.

WOMEN'S HEALTH INITIATIVE STUDY WRITING GROUP (1998). Design of the Women's Health Initiative clinical trial and observational study. *Control. Clin. Trials* **19** 61–109.

YU, R. (2020). Evaluating and improving a matched comparison of antidepressants and bone density. *Biometrics*. https://doi.org/10.1111/biom.13374

YU, R., SILBER, J. H. and ROSENBAUM, P. R. (2020). Matching methods for observational studies derived from large administrative databases. *Statist. Sci.* **35** 338–355. MR4148206 https://doi.org/10.1214/19-STS699

YU, R., SMALL, D. S. and ROSENBAUM, P. R. (2021). Supplement to "The information in covariate imbalance in studies of hormone replacement therapy." https://doi.org/10.1214/21-AOAS1448SUPP

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 2043–2060
https://doi.org/10.1214/21-AOAS1478

INFERRING FOOD INTAKE FROM MULTIPLE BIOMARKERS USING A LATENT VARIABLE MODEL

BY SILVIA D'ANGELO[1,*], LORRAINE BRENNAN[2] AND ISOBEL CLAIRE GORMLEY[1,†]

[1]*School of Mathematics and Statistics, Insight Centre for Data Analytics, University College Dublin,* [*]*silvia.dangelo@ucd.ie;* [†]*claire.gormley@ucd.ie*

[2]*School of Agriculture and Food Science, University College Dublin, lorraine.brennan@ucd.ie*

Metabolomic-based approaches have gained much attention in recent years, due to their promising potential to deliver objective tools for assessment of food intake. In particular, multiple biomarkers have emerged for single foods. However, there is a lack of statistical tools available for combining multiple biomarkers to quantitatively infer food intake. Furthermore, there is a paucity of approaches for estimating the uncertainty around biomarker-based inferred intake.

Here, to estimate the relationship between multiple metabolomic biomarkers and food intake in an intervention study conducted under the A-DIET research programme, a latent variable model, multiMarker, is proposed. The multiMarker model integrates factor analytic and mixture of experts models: the observed biomarker values are related to intake which is described as a continuous latent variable which follows a flexible mixture of experts model with Gaussian components. The multiMarker model also facilitates inference on the latent intake when only biomarker data are subsequently observed. A Bayesian hierarchical modelling framework provides flexibility to adapt to different biomarker distributions and facilitates inference of the latent intake along with its associated uncertainty.

Simulation studies are conducted to assess the performance of the multiMarker model, prior to its application to the motivating application of quantifying apple intake.

REFERENCES

AGRESTI, A. (1999). Modelling ordered categorical data: Recent advances and future challenges. *Stat. Med.* **18** 2191–2207.

BALDRICK, F. R., WOODSIDE, J. V., ELBORN, J. S., YOUNG, I. S. and MCKINLEY, M. C. (2011). Biomarkers of fruit and vegetable intake in human intervention studies: A systematic review. *Crit. Rev. Food Sci. Nutr.* **51** 795–815.

BARTHOLOMEW, D. J. and KNOTT, M. (1999). *Latent Variable Models and Factor Analysis*, 2nd ed. *Kendall's Library of Statistics* **7**. Edward Arnold, London; Oxford Univ. Press, New York. MR1711686

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

BINGHAM, S. A. (2002). Biomarkers in nutritional epidemiology. *Public Health Nutr.* **5** 821–827.

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

CAGNONE, S. and VIROLI, C. (2012). A factor mixture analysis model for multivariate binary data. *Stat. Model.* **12** 257–277. MR3179502 https://doi.org/10.1177/1471082X1101200303

CHEN, Y.-C., WANG, Y. S. and EROSHEVA, E. A. (2018). On the use of bootstrap with variational inference: Theory, interpretation, and a two-sample test example. *Ann. Appl. Stat.* **12** 846–876. MR3834288 https://doi.org/10.1214/18-AOAS1169

D'ANGELO, S., BRENNAN, L. and GORMLEY, I. C. (2020). multiMarker: Latent variable model to infer food intake from multiple biomarkers. R package version 1.0.1.

D'ANGELO, S., BRENNAN, L. and GORMLEY, I. C. (2021). Supplement to "Inferring food intake from multiple biomarkers using a latent variable model." https://doi.org/10.1214/21-AOAS1478SUPP

Key words and phrases. Latent variable models, factor analysis, mixture of experts, metabolomics, ordinal regression.

DRAGSTED, L. O. et al. (2018). Validation of biomarkers of food intake—critical assessment of candidate biomarkers. *Genes and Nutrition* **13**.

FRÜHWIRTH-SCHNATTER, S. and LOPES, H. F. (2018). Sparse Bayesian factor analysis when the number of factors is unknown. arXiv:1804.04231.

GALIMBERTI, G., MONTANARI, A. and VIROLI, C. (2009). Penalized factor mixture analysis for variable selection in clustered data. *Comput. Statist. Data Anal.* **53** 4301–4310. MR2744326 https://doi.org/10.1016/j.csda.2009.05.025

GAO, Q. et al. (2017). A scheme for a flexible classification of dietary and health biomarkers. *Genes and Nutrition* **12**.

GARCIA-ALOY, M., RABASSA, M., CASAS-AGUSTENCH, P., HIDALGO-LIBERONA, N., LLORACH, R. and ANDRES-LACUEVA, C. (2017). Novel strategies for improving dietary exposure assessment: Multiple-data fusion is a more accurate measure than the traditional single-biomarker approach. *Trends Food Sci. Technol.* **69** 220–229.

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

GÜRDENIZ, G. et al. (2016). Detecting beer intake by unique metabolite patterns. *J. Proteome Res.* **15** 4544–4556.

JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. and HINTON, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput.* **3** 79–87. https://doi.org/10.1162/neco.1991.3.1.79

KIPNIS, V. et al. (2002). Bias in dietary-report instruments and its implications for nutritional epidemiology. *Public Health Nutr.* **5** 915–923.

LIN, T.-I., MCLACHLAN, G. J. and LEE, S. X. (2016). Extending mixtures of factor models using the restricted multivariate skew-normal distribution. *J. Multivariate Anal.* **143** 398–413. MR3431441 https://doi.org/10.1016/j.jmva.2015.09.025

LLOYD, A. J., WILLIS, N. D., WILSON, T., ZUBAIR, H., CHAMBERS, E., GARCIA-PEREZ, I., XIE, L., TAILLIART, K., BECKMANN, M. et al. (2019). Addressing the pitfalls when designing intervention studies to discover and validate biomarkers of habitual dietary intake. *Metabolomics* **15**.

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

MCLACHLAN, G. J., BEAN, R. W. and JONES, L. B. (2007). Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution. *Comput. Statist. Data Anal.* **51** 5327–5338. MR2370874 https://doi.org/10.1016/j.csda.2006.09.015

MCNAMARA, A. E., COLLINS, C., SRI HARSHA, P. S. C., GONZÁLEZ-PEÑA, D., GIBBONS, H., MCNULTY, B. A., NUGENT, A. P., WALTON, J., FLYNN, A. et al. (2020). Metabolomic-based approach to identify biomarkers of apple intake. *Mol. Nutr. Food Res.*

MONTANARI, A. and VIROLI, C. (2010). Heteroscedastic factor mixture analysis. *Stat. Model.* **10** 441–460. MR2797648 https://doi.org/10.1177/1471082X0901000405

MORGAN, B. J. T. and SMITH, D. M. (1992). A note on Wadley's problem with overdispersion. *Appl. Stat.* **41** 287–497.

MURPHY, K., VIROLI, C. and GORMLEY, I. C. (2020). Infinite mixtures of infinite factor analysers. *Bayesian Anal.* **15** 937–963. MR4132655 https://doi.org/10.1214/19-BA1179

MURRAY, P. M., BROWNE, R. P. and MCNICHOLAS, P. D. (2014). Mixtures of skew-t factor analyzers. *Comput. Statist. Data Anal.* **77** 326–335. MR3210066 https://doi.org/10.1016/j.csda.2014.03.012

MUTHÉN, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika* **54** 557–585. MR1041525 https://doi.org/10.1007/BF02296397

PISON, G., ROUSSEEUW, P. J., FILZMOSER, P. and CROUX, C. (2003). Robust factor analysis. *J. Multivariate Anal.* **84** 145–172. MR1965827 https://doi.org/10.1016/S0047-259X(02)00007-6

R CORE TEAM (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

ROČKOVÁ, V. and GEORGE, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *J. Amer. Statist. Assoc.* **111** 1608–1622. MR3601721 https://doi.org/10.1080/01621459.2015.1100620

ROTHWELL, J. A. et al. (2014). New biomarkers of coffee consumption identified by the non-targeted metabolomic profiling of cohort study subjects. *PLoS ONE*.

SIDDIQUE, J., DANIELS, M. J., CARROLL, R. J., RAGHUNATHAN, T. E., STUART, E. A. and FREEDMAN, L. S. (2019). Measurement error correction and sensitivity analysis in longitudinal dietary intervention studies using an external validation study. *Biometrics* **75** 927–937. MR4012098 https://doi.org/10.1111/biom.13044

SUBAR, A. F., KIPNIS, V., TROIANO, R. P., MIDTHUNE, D., SCHOELLER, D. A., BINGHAM, S., SHARBAUGH, C. O., TRABULSI, J., RUNSWICK, S. et al. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The OPEN study. *Am. J. Epidemiol.* **158** 1–13.

VÁZQUEZ-MANJARREZ, N. et al. (2019). Discovery and validation of banana intake biomarkers using untargeted metabolomics in human intervention and cross-sectional studies. *J. Nutr.* **149** 1701–1713.

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 2061–2082
https://doi.org/10.1214/21-AOAS1467

MEDIATION ANALYSIS FOR ASSOCIATIONS OF CATEGORICAL VARIABLES: THE ROLE OF EDUCATION IN SOCIAL CLASS MOBILITY IN BRITAIN

BY JOUNI KUHA[1], ERZSÉBET BUKODI[2] AND JOHN H. GOLDTHORPE[3]

[1]*Department of Statistics, London School of Economics and Political Science, j.kuha@lse.ac.uk*
[2]*Department of Social Policy and Intervention and Nuffield College, University of Oxford, erzsebet.bukodi@spi.ox.ac.uk*
[3]*Nuffield College, University of Oxford, john.goldthorpe@nuffield.ox.ac.uk*

We analyse levels and trends of intergenerational social class mobility among three postwar birth cohorts in Britain and examine how much of the observed mobility or immobility in them could be accounted for by existing differences in educational attainment between people from different class backgrounds. We propose for this purpose a method which quantifies associations between categorical variables when we compare groups which differ only in the distribution of a mediating variable, such as education. This is analogous to estimation of indirect effects in causal mediation analysis but is here developed to define and estimate population associations of variables. We propose estimators for these associations which depend only on fitted values from models for the mediator and outcome variables, and propose variance estimators for them. The analysis shows that the part that differences in education play in intergenerational class mobility is by no means so dominant as has been supposed and that, while it varies with gender and with particular mobility transitions, it shows no tendency to change over time.

REFERENCES

AGRESTI, A. (2013). *Categorical Data Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR3087436

BARON, R. M. and KENNY, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51** 1173–1182.

BLALOCK, H. M. (1964). *Causal Inferences in Nonexperimental Research*. Univ. North Carolina Press, Chapel Hill, NC.

BLAU, P. M. and DUNCAN, O. D. (1967). *The American Occupational Structure*. Wiley, New York.

BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0996025 https://doi.org/10.1002/9781118619179

BREEN, R. and KARLSON, K. B. (2014). Education and social mobility: New analytical approaches. *Eur. Sociol. Rev.* **30** 107–118.

BREEN, R., KARLSON, K. B. and HOLM, A. (2013). Total, direct, and indirect effects in logit and probit models. *Sociol. Methods Res.* **42** 164–191. MR3190728 https://doi.org/10.1177/0049124113494572

BREEN, R. and MÜLLER, W., eds. (2020). *Education and Intergenerational Social Mobility in Europe and the United States*. Stanford Univ. Press, Stanford, CA.

BUKODI, E. and GOLDTHORPE, J. H. (2016). Educational attainment—relative or absolute—as a mediator of intergenerational class mobility in Britain. *Res. Soc. Stratif. Mobil.* **43** 5–15.

BUKODI, E. and GOLDTHORPE, J. H. (2018). *Social Mobility and Education in Britain*: *Research*, *Politics and Policy*. Cambridge Univ. Press, Cambridge.

BUKODI, E., GOLDTHORPE, J. H. and KUHA, J. (2017). The pattern of social fluidity within the British class structure: A topological model. *J. Roy. Statist. Soc. Ser. A* **180** 841–862. MR3660163 https://doi.org/10.1111/rssa.12234

BUKODI, E. and PASKOV, M. (2020). Intergenerational class mobility among men and women in Europe: Gender differences or gender similarities? *Eur. Sociol. Rev.* https://doi.org/10.1093/esr/jcaa001

Key words and phrases. Categorical data analysis, finite-population estimation, multinomial logistic models, path analysis.

BUKODI, E., PASKOV, M. and NOLAN, B. (2020). Intergenerational class mobility in Europe: A new account. *Soc. Forces* **98** 941–972.
BUKODI, E., GOLDTHORPE, J. H., WALLER, L. and KUHA, J. (2015). The mobility problem in Britain: New findings from the analysis of birth cohort data. *Br. J. Sociol.* **66** 93–117.
BUKODI, E., GOLDTHORPE, J. H., JOSHI, H. and WALLER, L. (2017). Why have relative rates of class mobility become more equal among women in Britain? *Br. J. Sociol.* **68** 512–532.
DAVIS, J. (1980). Contingency table analysis: Proportions and flow graphs. *Qual. Quant.* **14** 117–153.
DENIS, D. J. and LEGERSKI, J. (2006). Causal modeling and the origins of path analysis. *Theory Sci.* **7**.
DIDELEZ, V., DAWID, A. P. and GENELETTI, S. (2006). Direct and indirect effects of sequential treatments. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence* 138–146. Association for Uncertainty in Artificial Intelligence Press, Arlington, VA.
DUNCAN, O. D. and HODGE, R. W. (1963). Education and occupational mobility: A regression analysis. *Am. J. Sociol.* **68** 629–644.
ELLIOTT, J. and SHEPHERD, P. (2006). Cohort profile: 1970 British birth cohort (BCS70). *Int. J. Epidemiol.* **35** 836–843.
ERIKSON, R. and GOLDTHORPE, J. H. (1992). *The Constant Flux*: *A Study of Class Mobility in Industrial Societies*. Clarendon Press, Oxford.
GENELETTI, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 199–215. MR2325272 https://doi.org/10.1111/j.1467-9868.2007.00584.x
GOLDTHORPE, J. H. (2007). *On Sociology* **2**, 2nd ed. Stanford Univ. Press, Stanford, CA.
HECKMAN, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica* **46** 931–959. MR0483259 https://doi.org/10.2307/1909757
HELLEVIK, O. (1984). *Introduction to Causal Analysis*. George Allen & Unvin, London.
IMAI, K., KEELE, L. and TINGLEY, D. (2010). A general approach to causal mediation analysis. *Psychol. Methods* **15** 309–334.
IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—For Statistics, Social, and Biomedical Sciences*: *An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751
ISHIDA, H., MÜLLER, W. and RIDGE, J. (1995). Class origin, class destination and education: A cross-national study of industrial relations. *Am. J. Sociol.* **101** 145–193.
KUHA, J., BUKODI, E. and GOLDTHORPE, J. H. (2021a). Supplement to "Path analysis for discrete variables: The role of education in social mobility." https://doi.org/10.1214/21-AOAS1467SUPPA
KUHA, J., BUKODI, E. and GOLDTHORPE, J. H. (2021b). Computer code and pseudodata for "Path analysis for discrete variables: The role of education in social mobility." https://doi.org/10.1214/21-AOAS1467SUPPB
KUHA, J. and GOLDTHORPE, J. H. (2010). Path analysis for discrete variables: The role of education in social mobility. *J. Roy. Statist. Soc. Ser. A* **173** 351–369. MR2751881 https://doi.org/10.1111/j.1467-985X.2009.00620.x
LOEYS, T., MOERKERKE, B., DE SMET, O., BUYSSE, A., STEEN, J. and VANSTEELAND, S. (2013). Flexible mediation analysis in the presence of nonlinear relations: Beyond the mediation formula. *Multivar. Behav. Res.* **48** 871–894.
LÜTKEPOHL, H. (1996). *Handbook of Matrices*. Wiley, Chichester. MR1433592
MITNIK, P., CUMBERWORTH, E. and GRUSKY, D. (2016). Social mobility in a high-inequality regime. *Ann. Am. Acad. Polit. Soc. Sci.* **663** 140–184.
OFFICE FOR NATIONAL STATISTICS (2005). *The National Statistics Socio-Economic Classification*: *User Manual*. Palgrave–Macmillan, Basingstoke.
PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence* 411–420. Morgan Kaufmann, San Francisco, CA.
POWER, C. and ELLIOTT, J. (2006). Cohort profile: 1958 British birth cohort (National Child Development Study). *Int. J. Epidemiol.* **35** 34–41.
ROBINS, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems* (P. Green, N. Hjort and S. Richardson, eds.) 70–81. Oxford University Press, Oxford.
ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.
ROCKHILL, B., NEWMAN, B. and WEINBERG, C. (1998). Use and misuse of population attributable fractions. *Am. J. Publ. Health* **88** 15–19.
ROSE, D. and PEVALIN, D., eds. (2003). *A Researcher's Guide to the National Statistics Socio-Economic Classification*. Sage, London.
STATACORP (2017). Command *margins*. In *Stata* 15 *Base Reference Manual* Stata Press, College Station, TX.
TUKEY, J. W. (1954). Causation, regression and path analysis. In *Statistics and Mathematics in Biology* (O. Kempthorne, T. A. Bancroft, J. W. Gowen and J. L. Lush, eds.) 35–66. Hafner, New York.

VANDERWEELE, T. J. (2015). *Explanation in Causal Inference*: *Methods for Mediation and Interaction*. Oxford Univ. Press, New York.

VANDERWEELE, T. J. and ROBINSON, W. R. (2014). On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* **25** 473–484.

VANDERWEELE, T. J., VANSTEELANDT, S. and ROBINS, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator—outcome confounder. *Epidemiology* **25** 300–306.

WACHTER, K. W. (2014). *Essential Demographic Methods*. Harvard Univ. Press, Cambridge, MA.

WADSWORTH, M., KUH, D., RICHARDS, M. and HARDY, R. (2006). Cohort profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *Int. J. Epidemiol.* **35** 49–54. https://doi.org/10.1093/ije/dyi201

WINSHIP, C. and MARE, R. D. (1983). Structural equations and path analysis for discrete data. *Am. J. Sociol.* **89** 54–110.

WOLFLE, L. M. (2003). The introduction of path analysis to the social sciences, and some emergent themes: An annotated bibliography. *Struct. Equ. Model.* **10** 1–34. MR1951679 https://doi.org/10.1207/S15328007SEM1001_1

WRIGHT, S. (1921). Correlation and causation. *J. Agric. Res.* **20** 557–585.

XIE, Y. (1989). Structural equation models for ordinal variables: An analysis of occupational destination. *Sociol. Methods Res.* **17** 325–352.

The Annals of Applied Statistics
2021, Vol. 15, No. 4, 2083–2100
https://doi.org/10.1214/21-AOAS1464

PREDICTING COMPETITIONS BY COMBINING CONDITIONAL LOGISTIC REGRESSION AND SUBJECTIVE BAYES: AN ACADEMY AWARDS CASE STUDY

BY CHRISTOPHER T. FRANCK[1] AND CHRISTOPHER E. WILSON[2]

[1]*Department of Statistics, Virginia Tech, chfranck@vt.edu*
[2]*TIME, chris.wilson@time.com*

Predicting the outcome of elections, sporting events, entertainment awards and other competitions has long captured the human imagination. Such prediction is growing in sophistication in these areas, especially in the rapidly growing field of data-driven journalism intended for a general audience as the availability of historical information rapidly balloons. Providing statistical methodology to probabilistically predict competition outcomes faces two main challenges. First, a suitably general modeling approach is necessary to assign probabilities to competitors. Second, the modeling framework must be able to accommodate expert opinion which is usually available but difficult to fully encapsulate in typical data sets. We overcome these challenges with a combined conditional logistic regression/subjective Bayes approach. To illustrate the method, we reanalyze data from a recent Time.com piece in which the authors attempted to predict the 2019 Best Picture Academy Award winner using standard logistic regression. Toward engaging and educating a broad readership, we discuss strategies to deploy the proposed method via an online application.

REFERENCES

AGRESTI, A. (2013). *Categorical Data Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR3087436

BRESLOW, N. E., DAY, N. E., HALVORSEN, K. T., PRENTICE, R. L. and SABAI, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *Am. J. Epidemiol.* **108** 299–307.

BRODY, R. (2019). My 2019 Oscar predictions. Availabe at https://www.newyorker.com/culture/the-front-row/my-2019-oscars-predictions.

BYERS, D. (2012). Nate silver: One-term celebrity? Availabe at https://www.politico.com/blogs/media/2012/10/nate-silver-one-term-celebrity-147618.

CHERTOFF, E. (2012). Ironic etymology of the day: 'Pundit' comes from a sanskrit word for 'spiritual leader'. Availabe at https://www.theatlantic.com/national/archive/2012/08/ironic-etymology-of-the-day-pundit-comes-from-a-sanskrit-word-for-spiritual-leader/261493/.

COHN, N. (2017). A 2016 review: Why key state polls were wrong about Trump. Availabe at https://www.nytimes.com/2017/05/31/upshot/a-2016-review-why-key-state-polls-were-wrong-about-trump.html.

FRANCK, C. T. and WILSON, C. E. (2021). Supplement to "Predicting competitions by combining conditional logistic regression and subjective Bayes: An Academy Awards case study." https://doi.org/10.1214/21-AOAS1464SUPP

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GRAY, T. (2020). How does oscar voting work? Availabe at https://variety.com/feature/who-votes-on-oscars-academy-awards-how-voting-works-1203490944/.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417. With comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. MR1765176 https://doi.org/10.1214/ss/1009212519

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402 https://doi.org/10.1080/01621459.1995.10476572

Key words and phrases. Academy Awards, conditional logistic regression, prior elicitation, subjective Bayes.

KERR-DINEEN, L. (2017). Are 'win probabilities' useless? ESPN's Director of Sports Analytics explains why they're not. Availabe at https://ftw.usatoday.com/2017/02/super-bowl-espn-win-probability-atlanta-falcons-new-england-patriots-stats-tom-brady.

KING, A. (2019). Approximating the minds of 2019 Oscars voters using neural networks. Availabe at https://towardsdatascience.com/approximating-the-minds-of-2019-oscars-voters-using-neural-networks-b922f3d6864c.

LAWSON, R., COLLINS, K. A., ROBINSON, J., BUSIS, H. and HOGAN, M. (2019). 2019 Oscar predictions: Who V.F.'s experts expect to win this year. Availabe at https://www.vanityfair.com/hollywood/2019/02/oscars-2019-predictions-winners.

LENGEL, D. (2018). Has baseball analytics killed the art of hitting? Availabe at https://www.theguardian.com/sport/2018/oct/02/has-baseball-analytics-killed-the-art-of-hitting.

LI, Y. and CLYDE, M. A. (2018). Mixtures of g-priors in generalized linear models. *J. Amer. Statist. Assoc.* **113** 1828–1845. MR3902249 https://doi.org/10.1080/01621459.2018.1469992

LINZER, D. A. (2013). Dynamic Bayesian forecasting of presidential elections in the states. *J. Amer. Statist. Assoc.* **108** 124–134. MR3174607 https://doi.org/10.1080/01621459.2012.737735

LINZER, D. A. (2016). Forecasting the 2016 elections. Availabe at https://votamatic.org/forecasting-the-2016-elections/.

NGUYEN, T. D. (2015). 'Making big bucks' with a data-driven sports betting strategy. Availabe at https://towardsdatascience.com/making-big-bucks-with-a-data-driven-sports-betting-strategy-6c21a6869171.

O'HAGAN, A., BUCK, C., DANESHKHAH, A., EISER, J., GARTHWAITE, P., JENKINSON, D., OAKLEY, J. and RAKOW, T. (2006). *Uncertain Judgements*: *Eliciting Experts' Probabilities*. *Statistics in Practice*. Wiley, New York.

ROTHSCHILD, D. and WILSON, C. E. (2012). Obama poised to win 2012 election with 303 electoral votes: The Signal Forecast. Availabe at https://news.yahoo.com/blogs/signal/obama-poised-win-2012-election-303-electoral-votes-202543583.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAGwXJfWnpRN_80VPUzPTdm1twXK3fl6gd5uKd-hsmryJF30s3JKBfbW65WDZHNXQn3AEKqu9dLE0VsNI1DhZxZrWYotdtFPhSvSGVvSd0O4nLiKmoQN7TR3evSwdqcXp92F9Tdvt2OlN80QJ6H0qi_W4GfYYdSkoHPYuCGMvEAMf.

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

SILVER, N. (2012). Nov. 5: Late poll gains for obama leave romney with longer odds. Availabe at https://fivethirtyeight.com/features/nov-5-late-poll-gains-for-obama-leave-romney-with-longer-odds.

SILVER, N. (2016). Who will win the presidency? Availabe at https://projects.fivethirtyeight.com/2016-election-forecast/.

TANGO, T., LICHTMAN, M. and DOLPHIN, A. (2007). *The Book*: *Playing the Percentages in Baseball*. Potomac Books.

WILSON, C. E. (2019). How a harry potter quiz gave back to science. Availabe at https://time.com/5529413/harry-potter-quiz-sorting-hat/.

WILSON, C. E. and FRANCK, C. T. (2019). We taught a computer program to predict the Oscars. Here's the movie it says will win best picture. Availabe at https://time.com/5533849/oscars-academy-awards-prediction-best-picture/.

WILSON, C. E. and FRANCK, C. T. (2020). Our Oscars algorithm predicted the best picture winner. Tell us your guesses, too. Availabe at https://time.com/5779417/oscars-best-picture-prediction/.

YAO, Y., VEHTARI, A., SIMPSON, D. and GELMAN, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Anal.* **13** 917–1003. Including a rejoinder by the authors. MR3853125 https://doi.org/10.1214/17-BA1091

The Annals of Applied Statistics

Next Issues

Sparse matrix linear models for structured high-throughput data
JANE W. LIANG AND SAUNAK SEN

Modelled approximations to the ideal filter with application to GDP and its components
THOMAS M. TRIMBUR AND TUCKER MCELROY

Improving exoplanet detection power: Multivariate Gaussian process models for stellar activity . . . . . DAVID EDWARD JONES, DAVID C. STENNING, ERIC B. FORD, ROBERT L. WOLPERT, THOMAS J. LOREDO, CHRISTIAN GILBERTSON AND XAVIER DUMUSQUE

Bounding the local average treatment effect in an instrumental variable analysis of engagement with a mobile intervention . . . . . ANDREW JUSTIN SPIEKER, ROBERT GREEVY, LYNDSAY NELSON AND LINDSAY MAYBERRY

A functional-data approach to the Argo data
DREW YARGER, STILIAN STOEV AND TAILEN HSING

Model-based distance embedding with applications to chromosomal conformation biology
YUPING ZHANG, DISHENG MAO AND ZHENGQING OUYANG

A Bayesian model of dose-response for cancer drug studies
WESLEY TANSEY, CHRISTOPHER TOSH AND DAVID BLEI

Inference in Bayesian additive vector autoregressive tree models
FLORIAN HUBER AND LUCA ROSSINI

A flexible Bayesian framework to estimate age- and cause-specific child mortality over time from sample registration data . . . . . AUSTIN EDWARD SCHUMACHER, TYLER H. MCCORMICK, JON WAKEFIELD, YUE CHU, JAMIE PERIN, FRANCISCO VILLAVICENCIO, NOAH SIMON AND LI LIU

Robust causal inference for incremental return on ad spend with randomized paired geo experiments . . . . . AIYOU CHEN AND TIM AU

Bayesian nonparametric multivariate spatial mixture mixed effects models with application to American Community Survey special tabulations . . . . . RYAN JANICKI, ANDREW RAIM, SCOTT H. HOLAN AND JERRY MAPLES

Bidimensional linked matrix factorization for pan-omics pan-cancer analysis
ERIC F. LOCK, JUN YOUNG PARK AND KATHERINE A. HOADLEY

Multivariate mixed membership modeling: Inferring domain-specific risk profiles
MASSIMILIANO RUSSO

Fast inference for time-varying quantiles via flexible dynamic models with application to the characterization of atmospheric rivers . . . . . RAQUEL BARATA, RAQUEL PRADO AND BRUNO SANSO

Estimating the effectiveness of permanent price reductions for competing products using multivariate Bayesian structural time series models
FIAMMETTA MENCHETTI AND IAVOR BOJINOV

Bayesian adjustment for preferential testing in estimating infection fatality rates, as motivated by the COVID-19 pandemic . . . . . HARLAN CAMPBELL, PERRY DE VALPINE, LAUREN MAXWELL, VALENTIJN M. T. DE JONG, THOMAS P. A. DEBRAY, THOMAS JAENISCH AND PAUL GUSTAFSON

The assessment of replication success based on relative effect size
LEONHARD HELD, CHARLOTTE MICHELOUD AND SAMUEL PAWEL

Continued

The Annals of Applied Statistics

Next Issues—*Continued*

Subgroup-effects models for the analysis of personal treatment effects
LING ZHOU, SHIQUAN SUN, HAODA FU AND PETER X. K. SONG

Modeling non-stationary temperature maxima based on extremal dependence changing with event magnitude . . . . . PENG ZHONG, RAPHAEL HUSER AND THOMAS OPITZ

Sequential modeling, monitoring and forecasting of streaming web traffic data
KAORU IRIE, CHRISTOPHER GLYNN AND TEVFIK AKTEKIN

The role of intrinsic dimension in high-resolution player tracking data—Insights in basketball
EDGAR SANTOS FERNANDEZ, FRANCESCO DENTI, KERRIE MENGERSEN AND ANTONIETTA MIRA

Pre-electoral polls variability: A hierarchical Bayesian model to assess the role of house effects with application to Italian elections . . . . . DOMENICO DE STEFANO, FRANCESCO PAULI AND NICOLA TORELLI

Scalable changepoint and anomaly detection in cross-correlated data with an application to condition monitoring . . . . . MARTIN TVETEN, IDRIS ECKLEY AND PAUL FEARNHEAD

Detecting and modeling changes in a time series of proportions
THOMAS J. FISHER, JING ZHANG, STEPHEN COLEGATE AND MICHAEL J. VANNI

Prediction of hereditary cancers using neural networks
ZOE GUAN, GIOVANNI PARMIGIANI, DANIELLE BRAUN AND LORENZO TRIPPA

Identifying intergenerational patterns of correlated methylation sites
XICHEN MOU, HONGMEI ZHANG AND HASAN ARSHAD

Adaptive design for Gaussian process regression under censoring
JIALEI CHEN, SIMON MAK, V. ROSHAN JOSEPH AND CHUCK ZHANG

Ordinal probit functional outcome regression with application to computer-use behavior in rhesus monkeys . . . . . MARK J. MEYER, JEFFREY S. MORRIS, REGINA PAXTON GAZES AND BRENT A. COULL

In-game win probabilities for the National Rugby League . . . . . TIANYU GUAN, ROBERT NGUYEN, JIGUO CAO AND TIM SWARTZ

Composite mixture of log-linear models with application to psychiatric studies
EMANUELE ALIVERTI AND DAVID DUNSON

Partitioning around medoids clustering and random forest classification for GIS-informed imputation of fluoride concentration data . . . . . YU GU, JOHN PREISSER, DONGLIN ZENG, POOJAN SHRESTHA, MOLINA SHAH, MIGUEL SIMANCAS-PALLARES, JEANNIE GINNIS AND KIMON DIVARIS

Bayesian mitigation of spatial coarsening for a fairly flexible spatiotemporal Hawkes model
ANDREW JAMES HOLBROOK, XIANG JI AND MARC A. SUCHARD

Accounting for drop-out using inverse probability censoring weights in longitudinal clustered data with informative cluster size . . . . . AYA A. MITANI, ELIZABETH K. KAYE AND KERRIE P. NELSON

Inhomogeneous spatio-temporal point processes on linear networks for visitors' stops data
NICOLETTA D'ANGELO, GIADA ADELFIO, ANTONINO ABBRUZZO AND JORGE MATEU

Likelihood-based bacterial identification approach for bimicrobial mass spectrometry data
SO YOUNG RYU

Batch-sequential design and heteroskedastic surrogate modeling for delta smelt conservation
BOYA ZHANG, ROBERT B. GRAMACY, LEAH JOHNSON, KENNETH A. ROSE AND ERIC SMITH

Intensity estimation on geometric networks with penalized splines
MARC SCHNEBLE AND GÖRAN KAUERMANN

Sparse block signal detection and identification for shared cross-trait association analysis
JIANQIAO WANG, WANJIE WANG AND HONGZHE LI

Computationally efficient Bayesian unit-level models for non-Gaussian data under informative sampling . . . . . PAUL A. PARKER, SCOTT H. HOLAN AND RYAN JANICKI

Continued

The Annals of Applied Statistics

Next Issues—*Continued*

Approximate Bayesian inference for analysis of spatio-temporal flood frequency data
ÁRNI V. JÓHANNESSON, STEFAN SIEGERT, RAPHAEL HUSER, HAAKON BAKKA AND BIRGIR HRAFNKELSSON

Permutation tests under a rotating sampling plan with clustered data
JIAHUA CHEN, YUKUN LIU, CAROLYN TAYLOR AND JAMES ZIDAK

Inference for stochastic kinetic models from multiple data sources for joint estimation of infection dynamics from aggregate reports and virological data . . OKSANA A. CHKREBTII, YURY E. GARCIA, MARCOS A. CAPISTRAN AND DANIEL E. NOYOLA

Multi-state capture-recapture models for irregularly sampled data
SINA MEWS, ROLAND LANGROCK, RUTH KING AND NICOLA QUICK

Bayesian inverse reinforcement learning for collective animal movement
TORYN L. J. SCHAFER, CHRISTOPHER K. WIKLE AND MEVIN B. HOOTEN

A flexible sensitivity analysis approach for unmeasured confounding with multiple treatments and a binary outcome with applications to SEER-Medicare lung cancer data
LIANGYUAN HU, JUNGANG ZOU, CHENYANG GU, JIAYI JI, MICHAEL LOPEZ AND MINAL KALE

Robust Bayesian inference for big data: Combining sensor-based records with traditional survey data . . . . . ALI RAFEI, CAROL A. C. FLANNAGAN, BRADY T. WEST AND MICHAEL ELLIOTT

A sparse negative binomial classifier with covariate adjustment for RNA-seq data
TANBIN RAHMAN, HSIN-EN HUANG, YUJIA LI, AN-SHUN TAI, WEN-PING HSIEH AND GEORGE C. TSENG

Kernel machine and distributed lag models for assessing windows of susceptibility to environmental mixtures in children's health studies . . . . . ANDER WILSON, HSIAO-HSIEN LEON HSU, YUEH-HSIU MATHILDA CHIU, ROBERT O. WRIGHT, ROSALIND J. WRIGHT AND BRENT A. COULL

Contrastive latent variable modeling with application to case-control sequencing experiments
ANDREW JONES, F. WILLIAM TOWNES, DIDONG LI AND BARBARA E. ENGELHARDT

Detecting heterogeneous treatment effects with instrumental variables
MICHAEL WILLIAM JOHNSON, JIONGYI CAO AND HYUNSEUNG KANG

Statistical shape analysis of brain arterial networks (BAN)
XIAOYANG GUO, ADITI BASU BAL, TOM NEEDHAM AND ANUJ SRIVASTAVA

B-scaling: A novel nonparametric data fusion method
YIWEN LIU, XIAOXIAO SUN, WENXUAN ZHONG AND BING LI

Spatial-temporal-textual point processes for crime linkage detection
SHIXIANG ZHU AND YAO XIE

Markov-modulated Hawkes processes for modeling sporadic and bursty event occurrences in social interactions . . . . . JING WU, OWEN WARD, JAMES CURLEY AND TIAN ZHENG

Graph link prediction in computer networks using Poisson matrix factorisation
FRANCESCO SANNA PASSINO, MELISSA J. M. TURCOTTE AND NICHOLAS A. HEARD

Matrix completion methods for the total electron content video reconstruction
HU SUN, ZHIJUN HUA, JIAEN REN, SHASHA ZOU, YUEKAI SUN AND YANG CHEN

Conditional functional clustering for longitudinal data with heterogeneous nonlinear patterns
TIANHAO WANG, LEI YU, SUE E. LEURGANS, ROBERT S. WILSON, DAVID A. BENNETT AND PATRICIA A. BOYLE

Impact evaluation of the LAPD community safety partnership . . . . . SYDNEY KAHMANN, ERIN HARTMAN, JORJA LEAP AND P. JEFFREY BRANTINGHAM

Higher criticism for discriminating word-frequency tables and authorship attribution
ALON KIPNIS

The causal effect of a timeout at stopping an opposing run in the NBA
CONNOR PIERCE GIBBS, RYAN ELMORE AND BAILEY KATHRYN FOSDICK

Bayesian semiparametric long memory models for discretized event data
ANTIK CHAKRABORTY, OTSO OVASKAINEN AND DAVID DUNSON

Continued

The Annals of Applied Statistics

Next Issues—*Continued*

Co-clustering of multivariate functional data for the analysis of air pollution in the south of France . . . . . CHARLES BOUVEYRON, JULIEN JACQUES, AMANDINE SCHMUTZ, FANNY SIMOES AND SILVIA BOTTINI

Integrated Quantile RAnk Test (iQRAT) for gene-level associations in sequencing studies TIANYING WANG, IULIANA IONITA-LAZA AND YING WEI

A novel framework to estimate multi-dimensional minimum effective doses using asymmetric posterior gain and ε-tapering . . . . . YING KUEN CHEUNG, THEVAA CHANDERENG AND KEITH M. DIAZ

A Bayesian precision medicine framework for calibrating individualized therapeutic indices in cancer . . . . . ABHISEK SAHA, MIN JIN HA AND VEERA BALADANDAYUTHAPANI

Bayesian Local False Discovery Rate for sparse count data with application to the discovery of hotspots in protein domains . . . . . IRIS IVY M. GAURAN, JUNYONG PARK, ILIA RATTSEV, THOMAS A. PETERSON, MARICEL G. KANN AND DOHWAN PARK

Dirichlet-tree multinomial mixtures for clustering microbiome compositions JIALIANG MAO AND LI MA

Semiparametric point process modeling of blinking artifacts in PALM LOUIS GAMMELGAARD JENSEN, DAVID JOHN WILLIAMSON AND UTE HAHN

Semiparametric Bayesian forecasting of spatio-temporal earthquake occurrences GORDON J. ROSS AND ALEKSANDAR A. KOLEV

Improved inference on risk measures for univariate extremes LÉO RAYMOND BELZILE AND ANTHONY CHRISTOPHER DAVISON

A Bayesian hierarchical modeling approach to combining multiple data sources: A case study in size estimation . . . . . JACOB LEE PARSONS, XIAOYUE MAGGIE NIU AND LE BAO

Estimating mode effects from a sequential mixed-mode experiment using structural moment models . . . . . PAUL CLARKE AND YANCHUN BAO

Measuring performance for end-of-life care . . . . . SEBASTIEN HANEUSE, DEBORAH SCHRAG, FRANCESCA DOMINICI, SHARON-LISE NORMAND AND KYU HA LEE

Semiparametric multinomial mixed-effects models: A university students profiling tool CHIARA MASCI, FRANCESCA IEVA AND ANNA MARIA PAGANONI

Critical window variable selection for mixtures: Estimating the impact of multiple air pollutants on stillbirth . . . . . JOSHUA L. WARREN, HOWARD H. CHANG, LAUREN K. WARREN, MATTHEW J. STRICKLAND, LYNDSEY A. DARROW AND JAMES A. MULHOLLAND

High-resolution Bayesian mapping of landslide hazard with unobserved trigger event THOMAS OPITZ, HAAKON BAKKA, RAPHAËL HUSER AND LUIGI LOMBARDO

Bayesian functional registration of fMRI data GUOQING WANG, ABHIRUP DATTA AND MARTIN A. LINDQUIST

Joint integrative analysis of multiple data sources with correlated vector outcomes EMILY CHARLOTTE HECTOR AND PETER X.-K. SONG

Detection of two-way outliers in multivariate data and application to cheating detection in educational tests . . . . . YUNXIAO CHEN, YAN LU AND IRINI MOUSTAKI

Measurement error correction in particle tracking microrheology . . YUN LING, MARTIN LYSY, IAN SEIM, JAY NEWBY, DAVID HILL, JEREMY CRIBB AND M. GREGORY FOREST

Sensitivity analysis for evaluating principal surrogate endpoints relaxing the equal early clinical risk assumption . . . . . YING HUANG, YINGYING ZHUANG AND PETER GILBERT

Parameter calibration in wake effect simulation model with stochastic gradient descent and stratified sampling . . BINGJIE LIU, XUBO YUE, EUNSHIN BYON AND RAED AL KONTAR

Asymmetric tail dependence modeling, with application to cryptocurrency market data YAN GONG AND RAPHAEL HUSER

Analysis of presence-only data via exact Bayes, with model and effects identification GUIDO ALBERTI MOREIRA AND DANI GAMERMAN

Estimation of the marginal effect of antidepressants on body mass index under confounding and endogenous covariate-driven monitoring times . . . . . JANIE COULOMBE, ERICA E. M. MOODIE, ROBERT W. PLATT AND CHRISTEL RENOUX

Continued

The Annals of Applied Statistics

Next Issues—*Continued*

Estimating the stillbirth rate for 195 countries using a Bayesian sparse regression model with temporal smoothing . . ZHENGFAN WANG, MIRANDA J. FIX, LUCIA HUG, ANU MISHRA, DANZHEN YOU, HANNAH BLENCOWE, JON WAKEFIELD AND LEONTINE ALKEMA

Functional random effects modeling of brain shape and connectivity
EARDI LILA AND JOHN A. D. ASTON

Correction to: A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes . . . . . MARINA VANNUCCI

Hierarchical resampling for bagging in multi-study prediction with applications to human neurochemical sensing . . . . . GABRIEL LOEWINGER, PRASAD PATIL, KENNETH KISHIDA AND GIOVANNI PARMIGIANI

Large-scale multivariate sparse regression with applications to UK Biobank
JUNYANG QIAN, YOSUKE TANIGAWA, RUILIN LI, ROBERT TIBSHIRANI, MANUEL A. RIVAS AND TREVOR HASTIE