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

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

continued

Vol. 17, No. 1—March 2023

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*

Vol. 17, No. 1—March 2023

THE ANNALS OF APPLIED STATISTICS Vol. 17, No. 1, pp. 1–899 March 2023

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.

The Annals of Applied Statistics **[ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 17, Number 1, March 2023. 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
2023, Vol. 17, No. 1, 1–22
https://doi.org/10.1214/21-AOAS1583

FITTING STOCHASTIC EPIDEMIC MODELS TO GENE GENEALOGIES USING LINEAR NOISE APPROXIMATION

BY MINGWEI TANG[1,a], GYTIS DUDAS[2,3,b], TREVOR BEDFORD[3,c] AND VLADIMIR N. MININ[4,d]

[1]*Department of Statistics, University of Washington, Seattle,* [a]*tangmw1991@gmail.com*
[2]*Gothenburg Global Biodiversity Centre (GGBC),* [b]*gytisdudas@gmail.com*
[3]*Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center,* [c]*trevor@bedford.io*
[4]*Department of Statistics, University of California, Irvine,* [d]*vminin@uci.edu*

Phylodynamics is a set of population genetics tools that aim at reconstructing demographic history of a population based on molecular sequences of individuals sampled from the population of interest. One important task in phylodynamics is to estimate changes in (effective) population size. When applied to infectious disease sequences, such estimation of population size trajectories can provide information about changes in the number of infections. To model changes in the number of infected individuals, current phylodynamic methods use nonparametric approaches (e.g., Bayesian curve-fitting based on change-point models or Gaussian process priors), parametric approaches (e.g., based on differential equations), and stochastic modeling in conjunction with likelihood-free Bayesian methods. The first class of methods yields results that are hard to interpret epidemiologically. The second class of methods provides estimates of important epidemiological parameters, such as infection and removal/recovery rates, but ignores variation in the dynamics of infectious disease spread. The third class of methods is the most advantageous statistically but relies on computationally intensive particle filtering techniques that limits its applications. We propose a Bayesian model that combines phylodynamic inference and stochastic epidemic models and achieves computational tractability by using a linear noise approximation (LNA)—a technique that allows us to approximate probability densities of stochastic epidemic model trajectories. LNA opens the door for using modern Markov chain Monte Carlo tools to approximate the joint posterior distribution of the disease transmission parameters and of high dimensional vectors describing unobserved changes in the stochastic epidemic model compartment sizes (e.g., numbers of infectious and susceptible individuals). In a simulation study we show that our method can successfully recover parameters of stochastic epidemic models. We apply our estimation technique to Ebola genealogies estimated using viral genetic data from the 2014 epidemic in Sierra Leone and Liberia.

REFERENCES

ALTHAUS, C. L. (2014). Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. *PLoS Curr.* **6**. https://doi.org/10.1371/currents.outbreaks.91afb5e0f279e7f29e7056095255b288

ANDERSON, R. and MAY, R. (1992). *Infectious Diseases of Humans*: *Dynamics and Control* **28**. Wiley, New York.

BAILEY, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and Its Applications*, 2nd ed. Hafner Press, New York. MR0452809

BOUCKAERT, R., HELED, J., KÜHNERT, D., VAUGHAN, T., WU, C., XIE, D., SUCHARD, M., RAMBAUT, A. and DRUMMOND, A. (2014). BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput*. *Biol*. **10** 1–6.

Key words and phrases. Coalescent, susceptible-infectious-recovered model, state-space model, phylodynamics, Ebola virus.

BUCKINGHAM-JEFFERY, E., ISHAM, V. and HOUSE, T. (2018). Gaussian process approximations for fast inference from infectious disease data. *Math. Biosci.* **301** 111–120. MR3808358 https://doi.org/10.1016/j.mbs.2018.02.003

DEARLOVE, B. and WILSON, D. (2013). Coalescent inference for infectious disease: Meta-analysis of hepatitis C. *Philos. Trans. R. Soc. B* **368** 20120314.

DONNELLY, P. and TAVARE, S. (1995). Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29** 401–421.

DRUMMOND, A., NICHOLLS, G., RODRIGO, A. and SOLOMON, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161** 1307–1320.

DRUMMOND, A., RAMBAUT, A., SHAPIRO, B. and PYBUS, O. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22** 1185–1192.

DUDAS, G., CARVALHO, L., BEDFORD, T., TATEM, A., BAELE, G., FARIA, N., PARK, D., LADNER, J., ARIAS, A. et al. (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544** 309–315.

FEARNHEAD, P., GIAGOS, V. and SHERLOCK, C. (2014). Inference for reaction networks using the linear noise approximation. *Biometrics* **70** 457–466. MR3258050 https://doi.org/10.1111/biom.12152

FROST, S. D. and VOLZ, E. M. (2010). Viral phylodynamics and the search for an 'effective number of infections'. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **365** 1879–1890.

GIAGOS, V. (2010). Inference for Auto-Regulatory Genetic Networks Using Diffusion Process Approximations Ph.D. thesis Lancaster Univ.

GILL, M., LEMEY, P., FARIA, N., RAMBAUT, A., SHAPIRO, B. and SUCHARD, M. (2013). Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30** 713–724.

GILLESPIE, D. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81** 2340–2361.

GRENFELL, B., PYBUS, O., GOG, J., WOOD, J., DALY, J., MUMFORD, J. and HOLMES, E. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303** 327–332.

GRIFFITHS, R. and TAVARÉ, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **344** 403–410.

HÖHNA, S., LANDIS, M., HEATH, T., BOUSSAU, B., LARTILLOT, N., MOORE, B., HUELSENBECK, J. and RONQUIST, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* **65** 726–736.

JOMBART, T., CORI, A., DIDELOT, X., CAUCHEMEZ, S., FRASER, C. and FERGUSON, N. (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.* **10** e1003457. https://doi.org/10.1371/journal.pcbi.1003457

KARCHER, M., PALACIOS, J., BEDFORD, T., SUCHARD, M. and MININ, V. (2016). Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. *PLoS Comput. Biol.* **12** e1004789.

KEELING, M. and ROHANI, P. (2011). *Modeling Infectious Diseases in Humans and Animals*. Princeton Univ. Press, Princeton, NJ.

KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* **13** 235–248. MR0671034 https://doi.org/10.1016/0304-4149(82)90011-4

KLINKENBERG, D., BACKER, J. A., DIDELOT, X., COLIJN, C. and WALLINGA, J. (2017). Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput. Biol.* **13** e1005495. https://doi.org/10.1371/journal.pcbi.1005495

KOMOROWSKI, M., FINKENSTÄDT, B., HARPER, C. V. and RAND, D. A. (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinform.* **10** 343. https://doi.org/10.1186/1471-2105-10-343

KUHNER, M., YAMATO, J. and FELSENSTEIN, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149** 429–434.

KÜHNERT, D., STADLER, T., VAUGHAN, T. G. and DRUMMOND, A. J. (2014). Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. *J. R. Soc. Interface* **11** 20131106. https://doi.org/10.1098/rsif.2013.1106

KURTZ, T. G. (1970). Solutions of ordinary differential equations as limits of pure jump Markov processes. *J. Appl. Probab.* **7** 49–58. MR0254917 https://doi.org/10.2307/3212147

KURTZ, T. G. (1971). Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *J. Appl. Probab.* **8** 344–356. MR0287609 https://doi.org/10.1017/s002190020003535x

LEVENTHAL, G., GÜNTHARD, H., BONHOEFFER, S. and STADLER, T. (2013). Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. *Mol. Biol. Evol.* **31** 6–17.

MININ, V., BLOOMQUIST, E. and SUCHARD, M. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25** 1459–1471.

MÜLLER, N. F., RASMUSSEN, D. A. and STADLER, T. (2017). The structured coalescent and its approximations. *Mol. Biol. Evol.* **34** 2970–2981. https://doi.org/10.1093/molbev/msx186

MURRAY, I., ADAMS, R. and MACKAY, D. (2010). Elliptical slice sampling. In *AISTATS* **13** 541–548.

O'NEILL, P. and ROBERTS, G. (1999). Bayesian inference for partially observed stochastic epidemics. *J. Roy. Statist. Soc. Ser. A* **162** 121–129.

PALACIOS, J. A. and MININ, V. N. (2013). Gaussian process-based Bayesian nonparametric inference of population size trajectories from gene genealogies. *Biometrics* **69** 8–18. MR3058047 https://doi.org/10.1111/biom.12003

PAPASPILIOPOULOS, O., ROBERTS, G. O. and SKÖLD, M. (2007). A general framework for the parametrization of hierarchical models. *Statist. Sci.* **22** 59–73. MR2408661 https://doi.org/10.1214/088342307000000014

PYBUS, O., CHARLESTON, M., GUPTA, S., RAMBAUT, A., HOLMES, E. and HARVEY, P. (2001). The epidemic behavior of the hepatitis C virus. *Science* **292** 2323–2325.

RASMUSSEN, D. A., RATMANN, O. and KOELLE, K. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput. Biol.* **7** e1002136. MR2845064 https://doi.org/10.1371/journal.pcbi.1002136

RASMUSSEN, D. A., VOLZ, E. M. and KOELLE, K. (2014). Phylodynamic inference for structured epidemiological models. *PLoS Comput. Biol.* **10** e1003570. https://doi.org/10.1371/journal.pcbi.1003570

RUE, H. (2001). Fast sampling of Gaussian Markov random fields. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 325–338. MR1841418 https://doi.org/10.1111/1467-9868.00288

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields: Theory and Applications. Monographs on Statistics and Applied Probability* **104**. CRC Press/CRC, Boca Raton, FL. MR2130347 https://doi.org/10.1201/9780203492024

SCARPINO, S., IAMARINO, A., WELLS, C., YAMIN, D., NDEFFO-MBAH, M., WENZEL, N., FOX, S., NYENSWAH, T., ALTICE, F. et al. (2014). Epidemiological and viral genomic sequence analysis of the 2014 Ebola outbreak reveals clustered transmission. *Clin. Infect. Dis.* **60** 1079–1082.

SMITH, R. A., IONIDES, E. L. and KING, A. A. (2017). Infectious disease dynamics inferred from genetic data via sequential Monte Carlo. *Mol. Biol. Evol.* **34** 2065–2084. https://doi.org/10.1093/molbev/msx124

STADLER, T., KÜHNERT, D., BONHOEFFER, S. and DRUMMOND, A. (2013). Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci. USA* **110** 228–233.

STADLER, T., KÜHNERT, D., RASMUSSEN, D. and DU PLESSIS, L. (2014). Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. *PLoS Curr.* **6**.

SUCHARD, M. A., LEMEY, P., BAELE, G., AYRES, D. L., DRUMMOND, A. J. and RAMBAUT, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4** vey016. https://doi.org/10.1093/ve/vey016

TANG, M., DUDAS G., BEDFORD, T. and N. MININ, V. (2023). Supplement to "Fitting stochastic epidemic models to gene genealogies using linear noise approximation." https://doi.org/10.1214/21-AOAS1583SUPPA, https://doi.org/10.1214/21-AOAS1583SUPPB

TEAM, W. E. R. (2014). Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. *N. Engl. J. Med.* **371** 1481–1495.

TOWERS, S., PATTERSON-LOMBA, O. and CASTILLO-CHAVEZ, C. (2014). Temporal variations in the effective reproduction number of the 2014 West Africa Ebola outbreak. *PLoS Curr.* **6**. https://doi.org/10.1371/currents.outbreaks.9e4c4294ec8ce1adad283172b16bc908

VAN KAMPEN, N. and REINHARDT, W. (1981). *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam. MR0648937

VAUGHAN, T. G., LEVENTHAL, G. E., RASMUSSEN, D. A., DRUMMOND, A. J., WELCH, D. and STADLER, T. (2019). Estimating epidemic incidence and prevalence from genomic data. *Mol. Biol. Evol.* **36** 1804–1816. https://doi.org/10.1093/molbev/msz106

VOLZ, E. M. (2012). Complex population dynamics and the coalescent under neutrality. *Genetics* **190** 187–201. https://doi.org/10.1534/genetics.111.134627

VOLZ, E. M., KOELLE, K. and BEDFORD, T. (2013b). Viral phylodynamics. *PLoS Comput. Biol.* **9** e1002947. MR3048921 https://doi.org/10.1371/journal.pcbi.1002947

VOLZ, E. and POND, S. (2014). Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. *PLoS Curr.* **6**. https://doi.org/10.1371/currents.outbreaks.6f7025f1271821d4c815385b08f5f80e

VOLZ, E. and SIVERONI, I. (2018). Bayesian phylodynamic inference with complex models. *PLoS Comput. Biol.* **14** e1006546.

VOLZ, E., POND, S., WARD, M., BROWN, A. and FROST, S. (2009). Phylodynamics of infectious disease epidemics. *Genetics* **183** 1421–1430.

WALLACE, E. (2010). A simplified derivation of the linear noise approximation. Arxiv preprint. Available at arXiv:1004.4280.

WEARING, H. J., ROHANI, P. and KEELING, M. J. (2005). Appropriate models for the management of infectious diseases. *PLoS Med.* **2** e174. https://doi.org/10.1371/journal.pmed.0020174

WILKINSON, D. (2011). *Stochastic Modelling for Systems Biology*. CRC press, Boca Raton, FL.

WRIGHT, S. (1931). Evolution in Mendelian populations. *Genetics* **16** 97–159. https://doi.org/10.1093/genetics/16.2.97

XU, X., KYPRAIOS, T. and O'NEILL, P. D. (2016). Bayesian non-parametric inference for stochastic epidemic models using Gaussian processes. *Biostatistics* **17** 619–633. MR3604269 https://doi.org/10.1093/biostatistics/kxw011

YPMA, R. J. F., VAN BALLEGOOIJEN, W. M. and WALLINGA, J. (2013). Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* **195** 1055–1062.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2019). 2014–2016 Ebola outbreak in West Africa. https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/index.html. Last accessed: Oct, 09, 2022.

WORLD HEALTH ORGANIZATION (2016). Ebola data and statistics. http://apps.who.int/gho/data/node.ebola-sitrep.quick-downloads?lang=en. Last accessed: February 28, 2018.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 23–46
https://doi.org/10.1214/22-AOAS1613

SOCIAL DISTANCING AND COVID-19: RANDOMIZATION INFERENCE FOR A STRUCTURED DOSE-RESPONSE RELATIONSHIP

BY BO ZHANG[1,a], SIYU HENG[2,c], TING YE[3,d] AND DYLAN S. SMALL[1,b]

[1]*Department of Statistics and Data Science, The Wharton School, University of Pennsylvania,* [a]*bozhan@wharton.upenn.edu,* [b]*dsmall@wharton.upenn.edu*

[2]*Graduate Group in Applied Mathematics and Computational Science, School of Arts and Sciences, University of Pennsylvania,* [c]*siyuheng@sas.upenn.edu*

[3]*Department of Biostatistics, University of Washington,* [d]*tingye1@uw.edu*

Social distancing is widely acknowledged as an effective public health policy combating the novel coronavirus. But extreme forms of social distancing, like isolation and quarantine, have costs, and it is not clear how much social distancing is needed to achieve public health effects. In this article we develop a design-based framework to test the causal null hypothesis and make inference about the dose-response relationship between reduction in social mobility and COVID-19 related public health outcomes. We first discuss how to embed observational data with a time-independent, continuous treatment dose into an approximate randomized experiment and develop a randomization-based procedure that tests if a structured dose-response relationship fits the data. We then generalize the design and testing procedure to a longitudinal setting and apply them to investigate the effect of social distancing during the first phased reopening in the United States on public health outcomes using data compiled from Unacast™, the United States Census Bureau, and the County Health Rankings and Roadmaps Program. We rejected a primary analysis null hypothesis that stated the social distancing from April 27, 2020 to June 28, 2020, had no effect on the COVID-19-related death toll from June 29, 2020 to August 2, 2020 (p-value < 0.001), and found that it took more reduction in mobility to prevent exponential growth in case numbers for nonrural counties compared to rural counties.

REFERENCES

ACEMOGLU, D., CHERNOZHUKOV, V., WERNING, I. and WHINSTON, M. D. (2020). A multi-risk SIR model with optimally targeted lockdown. Technical Report. National Bureau of Economic Research.

ATALAN, A. (2020). Is the lockdown important to prevent the COVID-19 pandemic? Effects on psychology, environment and economy-perspective. *Ann. Med. Surg.* **56** 38–42.

ATHEY, S., ECKLES, D. and IMBENS, G. W. (2018). Exact p-values for network interference. *J. Amer. Statist. Assoc.* **113** 230–240. MR3803460 https://doi.org/10.1080/01621459.2016.1241178

BAIOCCHI, M., SMALL, D. S., LORCH, S. and ROSENBAUM, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Amer. Statist. Assoc.* **105** 1285–1296. MR2796550 https://doi.org/10.1198/jasa.2010.ap09490

BBC RADIO 4 (2020). Best of Today. https://www.bbc.co.uk/programmes/p08jn7g4.

BERGER, R. L. and BOOS, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* **89** 1012–1016. MR1294746

BIND, M.-A. C. and RUBIN, D. B. (2019). Bridging observational studies and randomized experiments by embedding the former in the latter. *Stat. Methods Med. Res.* **28** 1958–1978. MR3977087 https://doi.org/10.1177/0962280217740609

BONVINI, M., KENNEDY, E., VENTURA, V. and WASSERMAN, L. (2021). Causal inference in the time of Covid-19. Preprint arXiv:2103.04472.

BOWERS, J., FREDRICKSON, M. M. and PANAGOPOULOS, C. (2013). Reasoning about interference between units: A general framework. *Polit. Anal.* 97–124.

Key words and phrases. Causal inference, COVID-19, dose-response relationship, longitudinal studies, randomization inference, statistical matching.

BRAUER, F. and CASTILLO-CHAVEZ, C. (2012). *Mathematical Models in Population Biology and Epidemiology*, 2nd ed. *Texts in Applied Mathematics* **40**. Springer, New York. MR3024808 https://doi.org/10.1007/978-1-4614-1686-9

U. S. CENSUS BUREAU County Adjacency File.

CALABRESE, E. J. and BALDWIN, L. A. (2003). Hormesis: The dose–response revolution. *Annu. Rev. Pharmacol. Toxicol.* **43** 175–197.

CALLAGHAN, T., LUECK, J. A., TRUJILLO, K. L. and FERDINAND, A. O. (2021). Rural and urban differences in COVID-19 prevention behaviors. *J. Rural Health* **37** 287–295.

CHERNOZHUKOV, V., KASAHARA, H. and SCHRIMPF, P. (2021). Causal impact of masks, policies, behavior on early Covid-19 pandemic in the US. *J. Econometrics* **220** 23–62. MR4185124 https://doi.org/10.1016/j.jeconom.2020.09.003

DICKENS, B. L., KOO, J. R., WILDER-SMITH, A. and COOK, A. R. (2020). Institutional, not home-based, isolation could contain the COVID-19 outbreak. *Lancet* **395** 1541–1542.

DING, P., FELLER, A. and MIRATRIX, L. (2016). Randomization inference for treatment effect variation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 655–671. MR3506797 https://doi.org/10.1111/rssb.12124

FARRELLY, M. C., DAVIS, K. C., HAVILAND, M. L., MESSERI, P. and HEALTON, C. G. (2005). Evidence of a dose–response relationship between "truth" antismoking ads and youth smoking prevalence. *Am. J. Publ. Health* **95** 425–431.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, London and Edinburgh.

GELFAND, M. J., JACKSON, J. C., PAN, X., NAU, D., PIEPER, D., DENISON, E., DAGHER, M., VAN LANGE, P. A., CHIU, C.-Y. et al. (2021). The relationship between cultural tightness–looseness and COVID-19 cases and deaths: A global analysis. *Lancet Planetary Health* **5** e135–e144.

GORELL, J. M., RYBICKI, B. A., JOHNSON, C. C. and PETERSON, E. L. (1999). Smoking and Parkinson's disease: A dose–response relationship. *Neurology* **52** 115–115.

GROVER, S., SAHOO, S., MEHRA, A., AVASTHI, A., TRIPATHI, A., SUBRAMANYAN, A., PATTOJOSHI, A., RAO, G. P., SAHA, G. et al. (2020). Psychological impact of COVID-19 lockdown: An online survey from India. *Indian J. Psychiatry* **62** 354.

HENG, S., ZHANG, B., HAN, X., LORCH, S. A. and SMALL, D. S. (2019). Instrumental variables: To strengthen or not to strengthen? Preprint arXiv:1911.09171.

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* **15** 199–236.

HONG, G. and RAUDENBUSH, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J. Amer. Statist. Assoc.* **101** 901–910. MR2324091 https://doi.org/10.1198/016214506000000447

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

KOO, J. R., COOK, A. R., PARK, M., SUN, Y., SUN, H., LIM, J. T., TAM, C. and DICKENS, B. L. (2020). Interventions to mitigate early spread of SARS-CoV-2 in Singapore: A modelling study. *Lancet Infect. Dis.* **20** 678–688.

LAU, H., KHOSRAWIPOUR, V., KOCBACH, P., MIKOLAJCZYK, A., SCHUBERT, J., BANIA, J. and KHOSRAWIPOUR, T. (2020). The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China. *J. Travel Med.* **27**. https://doi.org/10.1093/jtm/taaa037

LAUER, S. A., GRANTZ, K. H., BI, Q., JONES, F. K., ZHENG, Q., MEREDITH, H. R., AZMAN, A. S., REICH, N. G. and LESSLER, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann. Intern. Med.* **172** 577–582.

LEWNARD, J. A. and LO, N. C. (2020). Scientific and ethical basis for social-distancing interventions against COVID-19. *Lancet Infect. Dis.* **20** 631–633. https://doi.org/10.1016/S1473-3099(20)30190-0

LU, B., ZANUTTO, E., HORNIK, R. and ROSENBAUM, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *J. Amer. Statist. Assoc.* **96** 1245–1253. MR1973668 https://doi.org/10.1198/016214501753381896

LU, B., GREEVY, R., XU, X. and BECK, C. (2011). Optimal nonbipartite matching and its statistical applications. *Amer. Statist.* **65** 21–30. MR2899649 https://doi.org/10.1198/tast.2011.08294

NOLEN, T. L. and HUDGENS, M. G. (2011). Randomization-based inference within principal strata. *J. Amer. Statist. Assoc.* **106** 581–593. MR2847972 https://doi.org/10.1198/jasa.2011.tm10356

WORLD HEALTH ORGANIZATION (2020). Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19).

REMINGTON, P. L., CATLIN, B. B. and GENNUSO, K. P. (2015). The county health rankings: Rationale and methods. *Popul. Health Metr.* **13** 11. https://doi.org/10.1186/s12963-015-0044-2

RICCIARDI, F., MATTEI, A. and MEALLI, F. (2020). Bayesian inference for sequential treatments under latent sequential ignorability. *J. Amer. Statist. Assoc.* **115** 1498–1517. MR4143481 https://doi.org/10.1080/01621459.2019.1623039

ROBINS, J. M. (1998). Marginal structural models. In 1997 *Proceedings of the Section on Bayesian Statistical Science*, American Statistical Association, Alexandria, VA, 1–10.

ROBINS, J., HERNÁN, M. A. and BABETTE, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550-560.

ROSENBAUM, P. R. (1989). Sensitivity analysis for matched observational studies with many ordered treatments. *Scand. J. Stat.* **16** 227–236. MR1040788

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. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Amer. Statist.* **59** 147–152. MR2133562 https://doi.org/10.1198/000313005X42831

ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *J. Amer. Statist. Assoc.* **102** 191–200. MR2345537 https://doi.org/10.1198/016214506000001112

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. *Springer Series in Statistics*. Springer, New York. MR2561612 https://doi.org/10.1007/978-1-4419-1213-8

RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *J. Amer. Statist. Assoc.* **81** 961–962.

RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071 https://doi.org/10.1198/016214504000001880

RUBIN, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat. Med.* **26** 20–36. MR2312697 https://doi.org/10.1002/sim.2739

SHERIDAN, A., ANDERSEN, A. L., HANSEN, E. T. and JOHANNESEN, N. (2020). Social distancing laws cause only small losses of economic activity during the COVID-19 pandemic in Scandinavia. *Proc. Natl. Acad. Sci. USA* **117** 20468–20473.

SJÖDIN, H., WILDER-SMITH, A., OSMAN, S., FAROOQ, Z. and ROCKLÖV, J. (2020). Only strict quarantine measures can curb the coronavirus disease (COVID-19) outbreak in Italy, 2020. *Eurosurveillance* **25** 2000280.

SOUCH, J. M. and COSSMAN, J. S. (2021). A commentary on rural-urban disparities in COVID-19 testing rates per 100,000 and risk factors. *J. Rural Health* **37** 188–190. https://doi.org/10.1111/jrh.12450

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313

TALLARIDA, R. J. and JACOB, L. S. (2012). *The Dose—Response Relation in Pharmacology*. Springer, Berlin.

TESTA, C. C., KRIEGER, N., CHEN, J. T. and HANAGE, W. P. (2020). Visualizing the lagged connection between COVID-19 cases and deaths in the United States: An animation using per capita state-level data (January 22, 2020–July 8, 2020). Technical Report. The Harvard Center for Population and Development Studies Munich.

THE NEW YORK TIMES (2020). Coronavirus (Covid-19) Data in the United States. https://github.com/nytimes/covid-19-data. Accessed: 2020-09-30.

UNNIKRISHNAN, C. (2020). Globally Coherent Weekly Periodicity in the Covid-19 Pandemic. medRxiv.

VENKATESH, A. and EDIRAPPULI, S. (2020). Social distancing in Covid-19: What are the mental health implications? *BMJ* **369** m1379. https://doi.org/10.1136/bmj.m1379

ZHANG, B., HENG, S., MACKAY, E. J. and YE, T. (2021). Bridging preference-based instrumental variable studies and cluster-randomized encouragement experiments: Study design, noncompliance, and average cluster effect ratio. *Biometrics* **in press**. https://doi.org/10.1111/biom.13500

ZHANG, B., HENG, S., YE, T. and SMALL, D. S. (2023). Supplement to "Social distancing and COVID-19: Randomization inference for a structured dose-response relationship." https://doi.org/10.1214/22-AOAS1613SUPPA, https://doi.org/10.1214/22-AOAS1613SUPPB, https://doi.org/10.1214/22-AOAS1613SUPPC

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 47–66
https://doi.org/10.1214/22-AOAS1615

CONTROL CHARTS FOR DYNAMIC PROCESS MONITORING WITH AN APPLICATION TO AIR POLLUTION SURVEILLANCE

BY XIULIN XIE[a] AND PEIHUA QIU[b]

Department of Biostatistics, University of Florida, [a]*xiulin.xie@ufl.edu,* [b]*pqiu@ufl.edu*

Air pollution is a major global public health risk factor. Among all air pollutants, $PM_{2.5}$ is especially harmful. It has been well demonstrated that chronic exposure to $PM_{2.5}$ can cause many health problems, including asthma, lung cancer and cardiovascular diseases. To tackle problems caused by air pollution, governments have put a huge amount of resources to improve air quality and reduce the impact of air pollution on public health. In this effort it is extremely important to develop an air pollution surveillance system to constantly monitor the air quality over time and to give a signal promptly once the air quality is found to deteriorate so that a timely government intervention can be implemented. To monitor a sequential process, a major statistical tool is the statistical process control (SPC) chart. However, traditional SPC charts are based on the assumptions that process observations at different time points are independent and identically distributed. These assumptions are rarely valid in environmental data because seasonality and serial correlation are common in such data. To overcome this difficulty, we suggest a new control chart in this paper, which can properly accommodate dynamic temporal pattern and serial correlation in a sequential process. Thus, it can be used for effective air pollution surveillance. This method is demonstrated by an application to monitor the daily average $PM_{2.5}$ levels in Beijing and shown to be effective and reliable in detecting the increase of $PM_{2.5}$ levels.

REFERENCES

AL-RASHED, A., AL-MUTAIRI, N. and ATTAR, M. A. (2019). Air pollution analysis in Kuwait using a statistical technique (CUSUM). *International Journal of Geosciences* **10** 254–294.

APLEY, D. W. and TSUNG, F. (2002). The autoregressive T^2 chart for monitoring univariate autocorrelated processes. *J. Qual. Technol.* **34** 80–96.

BARRATT, B., ATKINSON, R., ANDERSON, H. R., BEEVERS, S., KELLY, F., MUDWAY, I. and WILKINSON, P. (2007). Investigation into the use of the CUSUM technique in identifying changes in mean air pollution levels following introduction of a traffic management scheme. *Atmos. Environ.* **41** 1784–1791.

BOOGAARD, H., WALKER, K. and COHEN, A. J. (2019). Air pollution: The emergence of a major global health risk factor. *International Health* **11** 417–421.

CAPIZZI, G. and MASAROTTO, G. (2008). Practical design of generalized likelihood ratio control charts for autocorrelated data. *Technometrics* **50** 357–370. MR2528658 https://doi.org/10.1198/004017008000000280

CHAKRABORTI, S. and GRAHAM, M. A. (2019). Nonparametric (distribution-free) control charts: An updated overview and some results. *Quality Engineering* **31** 523–544.

CHALOULAKOU, A., KASSOMENOS, P., SPYRELLIS, N., DEMOKRITOU, P. and KOUTRAKIS, P. (2003). Measurements of PM10 and PM2.5 particle concentrations in Athens, Greece. *Atmos. Environ.* **37** 649–660.

CHATTERJEE, S. and QIU, P. (2009). Distribution-free cumulative sum control charts using bootstrap-based control limits. *Ann. Appl. Stat.* **3** 349–369. MR2668711 https://doi.org/10.1214/08-AOAS197

CHELANI, A. B. (2011). Change detection using CUSUM and modified CUSUM method in air pollutant concentrations at traffic site in Delhi. *Stoch. Environ. Res. Risk Assess.* **25** 827–834.

COHEN, A. J., BRAUER, M., BURNETT, R., ANDERSON, H. R., FROSTAD, J., ESTEP, K., BALAKRISHNAN, K., BRUNEKREEF, B., DANDONA, L. et al. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study 2015. *Lancet* **389** 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6

Key words and phrases. Data correlation, dynamic process monitoring, environmental protection, pollution surveillance, sequential learning, statistical process control.

CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatio-Temporal Data*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2848400

DE BRABANTER, K., DE BRABANTER, J., SUYKENS, J. A. K. and DE MOOR, B. (2011). Kernel regression in the presence of correlated errors. *J. Mach. Learn. Res.* **12** 1955–1976. MR2819023

EPSNEČNIKOV, V. A. (1969). Nonparametric estimation of a multidimensional probability density. *Theory Probab. Appl.* **14** 153–158.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. CRC Press, London. MR1383587

HE, S., JIANG, W. and DENG, H. (2018). A distance-based control chart for monitoring multivariate processes using support vector machines. *Ann. Oper. Res.* **263** 191–207. MR3775193 https://doi.org/10.1007/s10479-016-2186-4

HEALTH EFFECTS INSTITUTE (2019). *State of Global Air 2019*. Health Effects Institute, Boston, MA.

HIGHAM, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl.* **103** 103–118. MR0943997 https://doi.org/10.1016/0024-3795(88)90223-6

JACOB, D. J. and WINNER, D. A. (2009). Effect of climate change on air quality, atmospheric environment. *Atmos. Environ.* **43** 51–63.

LEE, H. C. and APLEY, D. W. (2011). Improved design of robust exponentially weighted moving average control charts for autocorrelated processes. *Qual. Reliab. Eng. Int.* **27** 337–352.

LI, J. and QIU, P. (2017). Construction of an efficient multivariate dynamic screening system. *Qual. Reliab. Eng. Int.* **30** 1969–1981.

LI, W., SHAO, L., WANG, W., LI, H., WANG, X., LI, Y., LI, W., JONES, T. and ZHANG, D. (2020). Air quality improvement in response to intensified control strategies in Beijing during 2013–2019. *Sci. Total Environ.* **744** 140776.

LIANG, X., ZOU, T., GUO, B., LI, S., ZHANG, H., ZHANG, S., HUANG, H. and CHEN, S. X. (2015). Assessing Beijing's PM2.5 pollution: Severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A* **471** 20150257.

LIANG, X., LI, S., ZHANG, S., HUANG, H. and CHEN, S. X. (2016). PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities. *J. Geophys. Res., Atmos.* **121** 220–236.

LIU, Y., ZHOU, Y. and LU, J. (2020). Exploring the relationship between air pollution and meteorological conditions in China under environmental governance. *Sci. Rep.* **10** 14518. https://doi.org/10.1038/s41598-020-71338-7

POPE, C. A., BURNETT, R. T., THURSTON, G. D., THUN, M. J., CALLE, E. E., KREWSKI, D. and GODLESKI, J. J. (2004). Cardiovascular mortality and long-term exposure to particulate air pollution: Epidemiological evidence of general pathophysiological pathways of disease. *Circulation* **109** 71–77.

QIU, P. (2014). *Introduction to Statistical Process Control*. Chapman Hall/CRC, Boca Raton, FL.

QIU, P. (2018). Some perspectives on nonparametric statistical process control. *J. Qual. Technol.* **50** 49–65.

QIU, P. and HAWKINS, D. (2003). A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions. *Statistician* **52** 151–164. MR1977257 https://doi.org/10.1111/1467-9884.00348

QIU, P. and XIANG, D. (2014). Univariate dynamic screening system: An approach for identifying individuals with irregular longitudinal behavior. *Technometrics* **56** 248–260. MR3207851 https://doi.org/10.1080/00401706.2013.822423

QIU, P. and XIE, X. (2021). Transparent sequential learning for statistical process control of serially correlated data. *Technometrics*. https://doi.org/10.1080/00401706.2021.1929493

SEAMAN, N. L. (2000). Meteorological modeling for air-quality assessments. *Atmos. Environ.* **34** 2231–2259.

SUKCHOTRAT, T., KIM, S. B. and TSUNG, F. (2010). One-class classification-based control charts for multivariate process monitoring. *IIE Trans.* **42** 107–120.

WU, W., JIN, Y. and CARLSTEN, C. (2018). Inflammatory health effects of indoor and outdoor particulate matter. *Journal of Allergy and Clinical Immunology* **141** 833–844.

XING, Y. F., XU, Y. H., SHI, M. H. and LIAN, Y. X. (2016). The impact of PM2.5 on the human respiratory system. *Journal of Thoracic Disease* **8** 69–74.

XUE, L. and QIU, P. (2021). A nonparametric CUSUM chart for monitoring multivariate serially correlated processes. *J. Qual. Technol.* **53** 396–409.

YANG, K. and QIU, P. (2020). Online sequential monitoring of spatio-temporal disease incidence rates. *IISE Trans.* **52** 1218–1233.

YOU, L. and QIU, P. (2019). Fast computing for dynamic screening systems when analyzing correlated data. *J. Stat. Comput. Simul.* **89** 379–394. MR3893031 https://doi.org/10.1080/00949655.2018.1552273

YOU, L. and QIU, P. (2020). An effective method for online disease risk monitoring. *Technometrics* **62** 249–264. MR4095749 https://doi.org/10.1080/00401706.2019.1625813

ZHANG, S., GUO, B., DONG, A., HE, J., XU, Z. and CHEN, S. X. (2017). Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A* **473** 20170457.

ZHAO, X., ZHANG, X., XU, X., XU, J., MENG, W. and PU, W. (2009). Seasonal and diurnal variations of ambient PM2.5 concentration in urban and rural environments in Beijing. *Atmos. Environ.* **43** 2893–2900.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 67–88
https://doi.org/10.1214/22-AOAS1616

COMPLEX DISCONTINUITY DESIGNS USING COVARIATES: IMPACT OF SCHOOL GRADE RETENTION ON LATER LIFE OUTCOMES IN CHILE

BY JUAN D. DÍAZ[1,a] AND JOSÉ R. ZUBIZARRETA[2,b]

[1]*Department of Management Control and Information Systems, University of Chile,* [a]*juadiaz@fen.uchile.cl*

[2]*Departments of Health Care Policy, Biostatistics, and Statistics, and CAUSALab, Harvard University,* [b]*zubizarreta@hcp.med.harvard.edu*

Regression discontinuity designs are extensively used for causal inference in observational studies. However, they are usually confined to settings with simple treatment rules and determined by a single running variable with a single cutoff. Motivated by the problem of estimating the impact of grade retention on educational and juvenile crime outcomes in Chile, we propose a framework and methods for complex discontinuity designs that encompass multiple treatment rules. In this framework the observed covariates play a central role for identification, estimation, and generalization of causal effects. Identification is nonparametric and relies on a local strong ignorability assumption. Estimation proceeds, as in any observational study, under strong ignorability, yet in a neighborhood of the cutoffs of the running variables. We discuss estimation approaches based on matching and weighting, including complementary regression modeling adjustments. We present assumptions for generalization, that is, for identification and estimation of average treatment effects for target populations. We also describe two approaches to select the neighborhood for analysis. We find that grade retention in Chile has a negative impact on future grade retention but is not associated with dropping out of school or committing a juvenile crime.

REFERENCES

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

ABADIE, A. and IMBENS, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *J. Bus. Econom. Statist.* **29** 1–11. MR2789386 https://doi.org/10.1198/jbes.2009.07333

ALMOND, D., DOYLE JR., J. J., KOWALSKI, A. E. and WILLIAMS, H. (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *Q. J. Econ.* **125** 591–634.

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

ANGRIST, J. D. and ROKKANEN, M. (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *J. Amer. Statist. Assoc.* **110** 1331–1344. MR3449027 https://doi.org/10.1080/01621459.2015.1012259

ATHEY, S., IMBENS, G. W. and WAGER, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 597–623. MR3849336 https://doi.org/10.1111/rssb.12268

BARRECA, A., GULDI, M., LINDO, J. and WADDELL, G. (2011). Saving babies? Revisiting the effect of very low birth weight classification. *Q. J. Econ.* **126** 2117–2123.

BARTALOTTI, O., BRUMMET, Q. and DIETERLE, S. (2021). A correction for regression discontinuity designs with group-specific mismeasurement of the running variable. *J. Bus. Econom. Statist.* **39** 833–848. MR4272939 https://doi.org/10.1080/07350015.2020.1737081

BATTISTIN, E. and RETTORE, E. (2008). Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. *J. Econometrics* **142** 715–730. MR2416826 https://doi.org/10.1016/j.jeconom.2007.05.006

BERTANHA, M. (2020). Regression discontinuity design with many thresholds. *J. Econometrics* **218** 216–241. MR4111751 https://doi.org/10.1016/j.jeconom.2019.09.010

Key words and phrases. Causal inference, observational studies, regression discontinuity design.

BRANSON, Z. and MEALLI, F. (2018). The local randomization framework for regression discontinuity designs: A review and some extensions. ArXiv arXiv–1810.

BRANSON, Z., RISCHARD, M., BORNN, L. and MIRATRIX, L. W. (2019). A nonparametric Bayesian methodology for regression discontinuity designs. *J. Statist. Plann. Inference* **202** 14–30. MR3926761 https://doi.org/10.1016/j.jspi.2019.01.003

CALONICO, S., CATTANEO, M. D. and TITIUNIK, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* **82** 2295–2326. MR3301169 https://doi.org/10.3982/ECTA11757

CARD, D. and SHORE-SHEPPARD, L. (2004). Using discontinuous eligibility rules to identify the effects of the federal medicaid expansions on low income children. *Rev. Econ. Stat.* **86** 752–766.

CATTANEO, M. D., FRANDSEN, B. R. and TITIUNIK, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *J. Causal Inference* **3** 1–24. MR4289424 https://doi.org/10.1515/jci-2013-0010

CATTANEO, M. D., IDROBO, N. and TITIUNIK, R. (2019). A practical introduction to regression discontinuity designs: Foundations.

CATTANEO, M. D., IDROBO, N. and TITIUNIK, R. (2020). *A Practical Introduction to Regression Discontinuity Designs: Extensions*. Cambridge Univ. Press, Cambridge.

CATTANEO, M. D. and VAZQUEZ-BARE, G. (2016). The choice of neighborhood in regression discontinuity designs. *Obs. Stud.* **2** 134–146.

CATTANEO, M. D., TITIUNIK, R., VAZQUEZ-BARE, G. and KEELE, L. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *J. Polit.* **78** 1229–1248.

CHAY, K., MCEWAN, P. and URQUIOLA, M. (2005). The central role of noise in evaluating interventions that use test scores to rank schools. *Am. Econ. Rev.* **95** 1237–1258.

CHOI, J.-Y. and LEE, M.-J. (2018). Regression discontinuity with multiple running variables allowing partial effects. *Polit. Anal.* **26** 258–274.

DE LA CUESTA, B. and IMAI, K. (2016). Misunderstandings about the regression discontinuity design in the study of close elections. *Annu. Rev. Pol. Sci.* **19** 375–396.

DE LA MAZA, D. (2012). The effect of medicaid eligibility on coverage, utilization, and children's health. *Health Econ.* **21** 1061–1079.

DÍAZ, J. D., GRAU, N., REYES, T. and RIVERA, J. (2021). The impact of primary school grade retention on juvenile crime. *Econ. Educ. Rev.*.

DÍAZ, J. D and ZUBIZARRETA, J. R (2023). Supplement to "Complex discontinuity designs using covariates: Impact of school grade retention on later life outcomes in Chile." https://doi.org/10.1214/22-AOAS1616SUPP

DONG, Y. (2015). Regression discontinuity applications with rounding errors in the running variable. *J. Appl. Econometrics* **30** 422–446. MR3340946 https://doi.org/10.1002/jae.2369

EGGER, P. and KOETHENBUERGER, M. (2010). Government spending and legislative organization: Quasi-experimental evidence from Germany. *Am. Econ. J. Appl. Econ.* **2** 200–212.

EREN, O., LOVENHEIM, M. and MOCAN, N. (2021). The effect of grade retention on adult crime: Evidence from a test-based promotion policy. *J. Labor Econ.*.

ERENA, O., DEPEW, B. and BARNES, S. (2017). Test-based promotion policies, dropping out, and juvenile crime. *J. Public Econ.* **153** 9–31.

FORASTIERE, L., MATTEI, A. and MEALLI, F. (2017). Selecting subpopulations for causal inference in regression discontinuity designs. Presented in the Workshop on The Regression Discontinuity Design: Methodological Issues and Applications in Economics, Statistics and Epidemiology.

GELMAN, A. and IMBENS, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *J. Bus. Econom. Statist.* **37** 447–456. MR3968384 https://doi.org/10.1080/07350015.2017.1366909

HAHN, J., TODD, P. and VAN DER KLAAUW, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69** 201–209.

HELLER, R., ROSENBAUM, P. R. and SMALL, D. S. (2010). Using the cross-match test to appraise covariate balance in matched pairs. *Amer. Statist.* **64** 299–309. MR2758561 https://doi.org/10.1198/tast.2010.09210

HIRSHBERG, D. A. and WAGER, S. (2021). Augmented minimax linear estimation. *Ann. Statist.* **49** 3206–3227. MR4352528 https://doi.org/10.1214/21-aos2080

IMBENS, G. W. and LEMIEUX, T. (2008). Regression discontinuity designs: A guide to practice. *J. Econometrics* **142** 615–635. MR2416821 https://doi.org/10.1016/j.jeconom.2007.05.001

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge Univ. Press, New York.

IMBENS, G. and WAGER, S. (2019). Optimized regression discontinuity designs. *Rev. Econ. Stat.* **101** 264–278.

JACOB, B. and LEFGREN, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Rev. Econ. Stat.* **86** 226–244.

KEELE, L. J. and TITIUNIK, R. (2015). Geographic boundaries as regression discontinuities. *Polit. Anal.* **23** 127–155.

KEELE, L., TITIUNIK, R. and ZUBIZARRETA, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *J. Roy. Statist. Soc. Ser. A* **178** 223–239. MR3291769 https://doi.org/10.1111/rssa.12056

KEELE, L., LORCH, S., PASSARELLA, M., SMALL, D. and TITIUNIK, R. (2017). An overview of geographically discontinuous treatment assignments with an application to children's health insurance. In *Regression Discontinuity Designs: Theory and Applications. Advances in Econometrics* **38** 147–94.

KERN, H. L., STUART, E. A., HILL, J. and GREEN, D. P. Assessing methods for generalizing experimental impact estimates to target populations. *J. Res. Educ. Eff.* **9** 103–127. https://doi.org/10.1080/19345747.2015.1060282

KOLESAR, M. and ROTHE, C. (2018). Inference in regression discontinuity designs with a discrete running variable. *Am. Econ. Rev.* **108** 2277–2304.

LEE, D. S. and CARD, D. (2008). Regression discontinuity inference with specification error. *J. Econometrics* **142** 655–674. MR2416823 https://doi.org/10.1016/j.jeconom.2007.05.003

LEE, D. S. and LEMIEUX, T. (2014). Regression discontinuity designs in social sciences. In *Regression Analysis and Causal Inference* (H. Best and C. Wolf, eds.), Sage.

LI, F., MATTEI, A. and MEALLI, F. (2015). Evaluating the causal effect of university grants on student dropout: Evidence from a regression discontinuity design using principal stratification. *Ann. Appl. Stat.* **9** 1906–1931. MR3456358 https://doi.org/10.1214/15-AOAS881

MANACORDA, M. (2012). The cost of grade retention. *Rev. Econ. Stat.* **94** 596–606.

MATSUDAIRA, J. D. (2008). Mandatory summer school and student achievement. *J. Econometrics* **142** 829–850. MR2416831 https://doi.org/10.1016/j.jeconom.2007.05.015

MATTEI, A. and MEALLI, F. (2016). Regression discontinuity designs as local randomized experiments. *Obs. Stud.* **66** 156–173.

PEI, Z. and SHEN, Y. (2017). *The Devil Is in the Tails: Regression Discontinuity Design with Measurement Error in the Assignment Variable*. Emerald.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

ROSENBAUM, P. R. (2002a). *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. (2002b). Attributing effects to treatment in matched observational studies. *J. Amer. Statist. Assoc.* **97** 183–192. MR1963391 https://doi.org/10.1198/016214502753479329

ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 515–530. MR2168202 https://doi.org/10.1111/j.1467-9868.2005.00513.x

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 SILBER, J. H. (2009). Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. *J. Amer. Statist. Assoc.* **104** 501–511. MR2751434 https://doi.org/10.1198/jasa.2009.0016

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–328.

RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–840. MR2516795 https://doi.org/10.1214/08-AOAS187

SALES, A. and HANSEN, B. B. (2020). Limitless regression discontinuity. *J. Educ. Behav. Stat.*. In press.

SCHWERDT, G., WEST, M. R. and WINTERS, M. A. (2017). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. *J. Public Econ.* **152** 154–169.

THISTLETHWAITE, D. and CAMPBELL, D. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *J. Educ. Psychol.* **51** 309–317.

WANG, Y. and ZUBIZARRETA, J. R. (2020). Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations. *Biometrika* **107** 93–105. MR4064142 https://doi.org/10.1093/biomet/asz050

WONG, V. C., STEINER, P. M. and COOK, T. D. (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *J. Educ. Behav. Stat.* **38** 107–141.

ZHAO, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Ann. Statist.* **47** 965–993. MR3909957 https://doi.org/10.1214/18-AOS1698

ZUBIZARRETA, J. R., PAREDES, R. D. and ROSENBAUM, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Ann*. *Appl*. *Stat*. **8** 204–231. MR3191988 https://doi.org/10.1214/13-AOAS713

ZUBIZARRETA, J. R., SMALL, D. S., GOYAL, N. K., LORCH, S. and ROSENBAUM, P. R. (2013). Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Ann*. *Appl*. *Stat*. **7** 25–50. MR3086409 https://doi.org/10.1214/12-AOAS582

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 89–107
https://doi.org/10.1214/22-AOAS1617

SOCIAL ORDER STATISTICS MODELS FOR RANKING DATA WITH ANALYSIS OF PREFERENCES IN SOCIAL NETWORKS

BY JIAQI GU[1,a] AND PHILIP L. H. YU[2,b]

[1]*Department of Statistics and Actuarial Science, University of Hong Kong,* [a]*u3005743@hku.hk*
[2]*Department of Mathematics and Information Technology, Education University of Hong Kong,* [b]*plhyu@eduhk.hk*

In the National Longitudinal Study of Adolescent to Adult Health (Add Health) in the United States during the 1994–95 school year, sampled high school students were asked to rank 17 romantic relationship activities in ideal chronological order and to nominate up to five male and five female best friends within the school. It is natural to see that the students' rank-order progression of social, romantic and sexual relationships may be influenced strongly by their peers or friends. So far, traditional ranking models do not account for such social network dependency. In this article we introduce a new class of models, called social order statistics (SOS) models, to learn ranking data in social networks. The new models combine the order statistics models and spatial autoregressive models to account for social dependencies among the individuals. A flexible formulation of weight matrices in the spatial model is adopted to provide diverse network effects among the individuals for different items. Efficient and scalable MCMC algorithms are developed to perform Bayesian inference in a parallel manner for large networks with even a few thousand nodes. Analysis of the Add Health dataset reveals that social network effects are different for students' preferences toward different activities in a relationship. In particular, students' preferences on romantic and sexual events generally have stronger peer effects than those on social events, and opinions of close friends of the same gender tend to have a larger impact on the preferences on sexual events.

REFERENCES

ALVO, M. and YU, P. L. H. (2014). *Statistical Methods for Ranking Data*. *Frontiers in Probability and the Statistical Sciences*. Springer, New York. MR3308929 https://doi.org/10.1007/978-1-4939-1471-5

BIERNACKI, C. and JACQUES, J. (2013). A generative model for rank data based on insertion sort algorithm. *Comput*. *Statist*. *Data Anal*. **58** 162–176. MR2997933 https://doi.org/10.1016/j.csda.2012.08.008

BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39** 324–345. MR0070925 https://doi.org/10.2307/2334029

CARON, F., TEH, Y. W. and MURPHY, T. B. (2014). Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. *Ann*. *Appl*. *Stat*. **8** 1145–1181. MR3262549 https://doi.org/10.1214/14-AOAS717

DEBARSY, N. and LESAGE, J. (2018). Flexible dependence modeling using convex combinations of different types of connectivity structures. *Regional Science and Urban Economics* **69** 48–68.

DENG, K., HAN, S., LI, K. J. and LIU, J. S. (2014). Bayesian aggregation of order-based rank data. *J*. *Amer*. *Statist*. *Assoc*. **109** 1023–1039. MR3265678 https://doi.org/10.1080/01621459.2013.878660

DIACONIS, P. (1989). A generalization of spectral analysis with application to ranked data. *Ann*. *Statist*. **17** 949–979. MR1015133 https://doi.org/10.1214/aos/1176347251

DOIGNON, J.-P., PEKEČ, A. and REGENWETTER, M. (2004). The repeated insertion model for rankings: Missing link between two subset choice models. *Psychometrika* **69** 33–54. MR2272438 https://doi.org/10.1007/BF02295838

FLIGNER, M. A. and VERDUCCI, J. S. (1988). Multistage ranking models. *J*. *Amer*. *Statist*. *Assoc*. **83** 892–901. MR0963820

GELMAN, A., ROBERTS, G. O. and GILKS, W. R. (1996). Efficient Metropolis jumping rules. In *Bayesian Statistics*, 5 (*Alicante*, 1994). *Oxford Sci*. *Publ*. 599–607. Oxford Univ. Press, New York. MR1425429

Key words and phrases. Ranking data, social network, MCMC, order statistics models.

GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, 4 (*Peñíscola*, 1991) J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 169–193. Oxford Univ. Press, New York. MR1380276

GORMLEY, I. C. and MURPHY, T. B. (2008). 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

GORMLEY, I. C. and MURPHY, T. B. (2009). A grade of membership model for rank data. *Bayesian Anal.* **4** 265–295. MR2507364 https://doi.org/10.1214/09-BA410

GORMLEY, I. C. and MURPHY, T. B. (2010). Clustering ranked preference data using sociodemographic covariates. In *Choice Modelling: The State-of-the-Art and the State-of-Practice* 543–569. Emerald Group Publishing Limited.

GU, J. and YU, P. L. H. (2023). Supplement to "Social order statistics models for ranking data with analysis of preferences in social networks. https://doi.org/10.1214/22-AOAS1617SUPP

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

HALL, P. and MILLER, H. (2010). Modeling the variability of rankings. *Ann. Statist.* **38** 2652–2677. MR2722452 https://doi.org/10.1214/10-AOS794

HAUSMAN, J. A. and RUUD, P. A. (1987). Specifying and testing econometric models for rank-ordered data. *J. Econometrics* **34** 83–104. MR0887389 https://doi.org/10.1016/0304-4076(87)90068-6

HOLLOWAY, G., SHANKAR, B. and RAHMANB, S. (2002). Bayesian spatial probit estimation: A primer and an application to HYV rice adoption. *Agric. Econ.* **27** 383–402.

HUANG, T.-K., WENG, R. C. and LIN, C.-J. (2006). Generalized Bradley–Terry models and multi-class probability estimates. *J. Mach. Learn. Res.* **7** 85–115. MR2274363

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

JORDAN, M. I. and JACOBS, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Comput.* **6** 181–214.

LEE, P. H. and YU, P. L. H. (2012). Mixtures of weighted distance-based models for ranking data with applications in political studies. *Comput. Statist. Data Anal.* **56** 2486–2500. MR2910064 https://doi.org/10.1016/j.csda.2012.02.002

LESAGE, J. P. (2008). An introduction to spatial econometrics. *Revue D'économie Industrielle* **123** 19–44.

LESAGE, J. and PACE, R. K. (2009). *Introduction to Spatial Econometrics. Statistics: Textbooks and Monographs*. CRC Press, Boca Raton, FL. MR2485048 https://doi.org/10.1201/9781420064254

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40. MR1279653 https://doi.org/10.1093/biomet/81.1.27

LUCE, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York. MR0108411

MALLOWS, C. L. (1957). Non-null ranking models. I. *Biometrika* **44** 114–130. MR0087267 https://doi.org/10.1093/biomet/44.1-2.114

MARDEN, J. I. (1995). *Analyzing and Modeling Rank Data. Monographs on Statistics and Applied Probability* **64**. CRC Press, London. MR1346107

MEILĂ, M. and BAO, L. (2008). Estimation and clustering with infinite rankings. 393–402.

MEILĂ, M. and BAO, L. (2010). An exponential model for infinite rankings. *J. Mach. Learn. Res.* **11** 3481–3518. MR2756191

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

NEISWANGER, W., WANG, C. and XING, E. P. (2014). Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence. UAI'*14 623–632. AUAI Press, Arlington, VA.

NEWMAN, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577–8582.

NG, A., JORDAN, M. and WEISS, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* (T. Dietterich, S. Becker and Z. Ghahramani, eds.) **14**. MIT Press, Cambridge.

PAKMAN, A. and PANINSKI, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *J. Comput. Graph. Statist.* **23** 518–542. MR3215823 https://doi.org/10.1080/10618600.2013.788448

PLACKETT, R. L. (1975). The analysis of permutations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **24** 193–202. MR0391338 https://doi.org/10.2307/2346567

RAJU, E., HAMEED, M. A. and SRAVANTHI, K. (2015). Detecting communities in social networks using unnormalized spectral clustering incorporated with bisecting K-means. In 2015 *IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)* IEEE, New York.

RITTER, C. and TANNER, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs sampler. *J. Amer. Statist. Assoc.* **87** 861–868.

SHI, J. and MALIK, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 888–905.

SPIELMAN, D. A. and TENG, S.-H. (2007). Spectral partitioning works: Planar graphs and finite element meshes. *Linear Algebra Appl.* **421** 284–305. MR2294342 https://doi.org/10.1016/j.laa.2006.07.020

TERENIN, A., SIMPSON, D. and DRAPER, D. (2020). Asynchronous Gibbs sampling (S. Chiappa and R. Calandra, eds.). *Proceedings of Machine Learning Research* **108** 144–154. PMLR, Online.

THURSTONE, L. L. (1927). A law of comparative judgment. *Psychological Review* **34** 273.

VITELLI, V., SØRENSEN, Ø., CRISPINO, M., FRIGESSI, A. and ARJAS, E. (2017). Probabilistic preference learning with the Mallows rank model. *J. Mach. Learn. Res.* **18** Paper No. 158. MR3813807

WANG, H., IGLESIAS, E. M. and WOOLDRIDGE, J. M. (2013). Partial maximum likelihood estimation of spatial probit models. *J. Econometrics* **172** 77–89. MR2997131 https://doi.org/10.1016/j.jeconom.2012.08.005

WANG, Y. S., MATSUEDA, R. L. and EROSHEVA, E. A. (2017). A variational EM method for mixed membership models with multivariate rank data: An analysis of public policy preferences. *Ann. Appl. Stat.* **11** 1452–1480. MR3709566 https://doi.org/10.1214/17-AOAS1034

WHITE, S. and SMYTH, P. (2005). A spectral clustering approach to finding communities in graphs. In *Proceedings of the* 2005 *SIAM International Conference on Data Mining* SIAM, Philadelphia.

XU, Y., ZHUANG, Z., LI, W. and ZHOU, X. (2018). Effective community division based on improved spectral clustering. *Neurocomputing* **279** 54–62.

YU, P. L. H. (2000). Bayesian analysis of order-statistics models for ranking data. *Psychometrika* **65** 281–299.

YU, P. L. H., GU, J. and XU, H. (2019). Analysis of ranking data. *Wiley Interdiscip. Rev.: Comput. Stat.* **11** e1483. MR4021451 https://doi.org/10.1002/wics.1483

YU, P. L. H., LAM, K. F. and LO, S. M. (2005). Factor analysis for ranked data with application to a job selection attitude survey. *J. Roy. Statist. Soc. Ser. A* **168** 583–597. MR2146411 https://doi.org/10.1111/j.1467-985X.2005.00363.x

YU, P. L. H. and XU, H. (2019). Rank aggregation using latent-scale distance-based models. *Stat. Comput.* **29** 335–349. MR3914624 https://doi.org/10.1007/s11222-018-9811-9

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 108–129
https://doi.org/10.1214/22-AOAS1618

PROBABILISTIC HIV RECENCY CLASSIFICATION—A LOGISTIC REGRESSION WITHOUT LABELED INDIVIDUAL LEVEL TRAINING DATA

BY BEN SHENG[1,a], CHANGCHENG LI[2,d], LE BAO[1,b] AND RUNZE LI[1,c]

[1]*Department of Statistics, Penn State University,* [a]*bxs416@psu.edu,* [b]*lebao@psu.edu,* [c]*rzli@psu.edu*
[2]*School of Mathematical Sciences, Dalian University of Technology,* [d]*lichangcheng@dlut.edu.cn*

Accurate HIV incidence estimation, based on individual recent infection status (recent vs long-term infection), is important for monitoring the epidemic, targeting interventions to those at greatest risk of new infection, and evaluating existing programs of prevention and treatment. Starting from 2015, the population-based HIV impact assessment (PHIA) individual-level surveys are implemented in the most-affected countries in sub-Saharan Africa. PHIA is a nationally-representative HIV-focused survey that combines household visits with key questions and cutting-edge technologies, such as biomarker tests for HIV antibody and HIV viral load which offer the unique opportunity of distinguishing between recent infection and long-term infection, and providing relevant HIV information by age, gender, and location. In this article we propose a semisupervised logistic regression model for estimating individual level HIV recency status. It incorporates information from multiple data sources—the PHIA survey, where the true HIV recency status is unknown, and the cohort studies provided in the literature where the relationship between HIV recency status and the covariates are presented in the form of a contingency table. It also utilizes the national level HIV incidence estimates from the epidemiology model. Applying the proposed model to Malawi PHIA data, we demonstrate that our approach is more accurate for the individual level estimation and more appropriate for estimating HIV recency rates at aggregated levels than the current practice—the binary classification tree (BCT).

REFERENCES

ICAP AT COLUMBIA UNIVERSITY (2020). Surveys—PHIA Project https://phia.icap.columbia.edu/surveys/ (accessed 1 February 2021).

BAO, L. (2012). A new infectious disease model for estimating and projecting HIV/AIDS epidemics. *Sex. Transm. Infect.* **88** i58–i64.

BAO, L. and RAFTERY, A. E. (2010). A stochastic infection rate model for estimating and projecting national HIV prevalence rates. *Sex. Transm. Infect.* **86** ii93–ii99. https://doi.org/10.1136/sti.2010.044529

BAO, L., YE, J. and HALLETT, T. B. (2014). Incorporating incidence information within the UNAIDS estimation and projection package framework: A study based on simulated incidence assay data. *AIDS* **28** S515.

BAO, L., SALOMON, J. A., BROWN, T., RAFTERY, A. E. and HOGAN, D. R. (2012). Modelling national HIV/AIDS epidemics: Revised approach in the UNAIDS estimation and projection package 2011. *Sex. Transm. Infect.* **88** i3–i10.

BARAK, B., CHAUDHURI, K., DWORK, C., KALE, S., MCSHERRY, F. and TALWAR, K. (2007). Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* 273–282.

BRESLOW, N. E. (1996). Statistics in epidemiology: The case-control study. *J. Amer. Statist. Assoc.* **91** 14–28. MR1394064 https://doi.org/10.2307/2291379

BROWN, T., BAO, L., RAFTERY, A. E., SALOMON, J. A., BAGGALEY, R. F., STOVER, J. and GERLAND, P. (2010). EPP 2009: Bringing the UNAIDS estimation and projection package into the ART era (supported by NICHD grant HD054511). *Sex. Transm. Infect.* **86** ii3–ii10.

Key words and phrases. Weakly supervised learning, logistic regression, contingency table, HIV recency, HIV incidence.

BROWN, T., BAO, L., EATON, J. W., HOGAN, D. R., MAHY, M., MARSH, K., MATHERS, B. M. and PUCKETT, R. (2014). Improvements in prevalence trend fitting and incidence estimation in EPP 2013. *AIDS* **28** S415–S425. https://doi.org/10.1097/QAD.0000000000000454

CARBONNEAU, M.-A., CHEPLYGINA, V., GRANGER, E. and GAGNON, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit*. **77** 329–353.

CORI, A., PICKLES, M., VAN SIGHEM, A., GRAS, L., BEZEMER, D., REISS, P. and FRASER, C. (2015). CD4+ cell dynamics in untreated HIV-1 infection: Overall rates, and effects of age, viral load, sex and calendar time. *AIDS* **29** 2435.

DUONG, Y. T., KASSANJEE, R., WELTE, A., MORGAN, M., DE, A., DOBBS, T., ROTTINGHAUS, E., NKENGASONG, J., CURLIN, M. E. et al. (2015). Recalibration of the limiting antigen avidity EIA to determine mean duration of recent infection in divergent HIV-1 subtypes. *PLoS ONE* **10** e0114947. https://doi.org/10.1371/journal.pone.0114947

DZIAK, J. J., COFFMAN, D. L., LANZA, S. T., LI, R. and JERMIIN, L. S. (2020). Sensitivity and specificity of information criteria. *Brief. Bioinform*. **21** 553–565.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903 https://doi.org/10.1007/978-1-4899-4541-9

ELKAN, C. and NOTO, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 213–220.

ESBER, A. L., COAKLEY, P., AKE, J. A., BAHEMANA, E., ADAMU, Y., KIWEEWA, F., MASWAI, J., OWUOTH, J., ROBB, M. L. et al. (2020). Decreasing time to antiretroviral therapy initiation after HIV diagnosis in a clinic-based observational cohort study in four African countries. *J. Int. AIDS Soc*. **23** e25446.

FELLOWS, I. E., SHIRAISHI, R. W., CHERUTICH, P., ACHIA, T., YOUNG, P. W. and KIM, A. A. (2020). A new method for estimating HIV incidence from a single cross-sectional survey. *PLoS ONE* **15** e0237221.

FIENBERG, S. E., RINALDO, A. and YANG, X. (2010). Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *International Conference on Privacy in Statistical Databases* 187–199. Springer, Berlin.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2017). New PHIA Survey Data Show Critical Progress Towards Global HIV Targets https://www.cdc.gov/globalhivtb/who-we-are/events/world-aids-day/phia-surveys.html (accessed 30 September 2020).

GIGUÈRE, K., EATON, J. W., MARSH, K., JOHNSON, L. F., JOHNSON, C. C., EHUI, E., JAHN, A., WANYEKI, I., MBOFANA, F. et al. (2021). Trends in knowledge of HIV status and efficiency of HIV testing services in sub-Saharan Africa, 2000–20: A modelling study using survey and HIV testing programme data. *The Lancet HIV* **6** e284–e293.

HADER, S. (2016). Innovation to Drive Impact: Reaching the Hardest to Reach. https://blogs.cdc.gov/global/2016/12/01/innovation-to-drive-impact-reaching-the-hardest-to-reach/ (accessed 30 September 2020).

HALLETT, T. B. (2011). Estimating the HIV incidence rate–recent and future developments. *Current Opinion in HIV and AIDS* **6** 102.

HOGAN, D. R., ZASLAVSKY, A. M., HAMMITT, J. K. and SALOMON, J. A. (2010). Flexible epidemiological model for estimates and short-term projections in generalised HIV/AIDS epidemics. *Sex. Transm. Infect*. **86** ii84–ii92. https://doi.org/10.1136/sti.2010.045104

JONES, J., SULLIVAN, P. S. and CURRAN, J. W. (2019). Progress in the HIV epidemic: Identifying goals and measuring success. *PLoS Med*. **16** e1002729.

JUSTMAN, J. E., MUGURUNGI, O. and EL-SADR, W. M. (2018). HIV population surveys—bringing precision to the global response. *N. Engl. J. Med*. **378** 1859–1861. https://doi.org/10.1056/NEJMp1801934

KAHN, J. O. and WALKER, B. D. (1998). Acute human immunodeficiency virus type 1 infection. *N. Engl. J. Med*. **339** 33–39.

KASSANJEE, R., MCWALTER, T. A., BÄRNIGHAUSEN, T. and WELTE, A. (2012). A new general biomarker-based incidence estimator. *Epidemiology* **23** 721.

KASSANJEE, R., PILCHER, C. D., BUSCH, M. P., MURPHY, G., FACENTE, S. N., KEATING, S. M., MCKINNEY, E., MARSON, K., PRICE, M. A. et al. (2016). Viral load criteria and threshold optimization to improve HIV incidence assay characteristics-a CEPHIA analysis. *AIDS* **30** 2361–2371.

LAVREYS, L., BAETEN, J. M., OVERBAUGH, J., PANTELEEFF, D. D., CHOHAN, B. H., RICHARDSON, B. A., MANDALIYA, K., NDINYA-ACHOLA, J. O. and KREISS, J. K. (2002). Virus load during primary human immunodeficiency virus (HIV) type 1 infection is related to the severity of acute HIV illness in Kenyan women. *Clin. Infect. Dis*. **35** 77–81.

LEE, W. S. and LIU, B. (2003). Learning with positive and unlabeled examples using weighted logistic regression. In *ICML* **3** 448–455.

MALAWI MINISTRY OF HEALTH (2018). Malawi Population-based HIV Impact Assessment (MPHIA) 2015-16: Final Report, Lilongwe.

MUSICANT, D. R., CHRISTENSEN, J. M. and OLSON, J. F. (2007). Supervised learning by training on aggregate outputs. In *Seventh IEEE International Conference on Data Mining* (*ICDM* 2007) 252–261. https://doi.org/10.1109/ICDM.2007.50

PANTAZIS, N., THOMADAKIS, C., DEL AMO, J., ALVAREZ-DEL ARCO, D., BURNS, F. M., FAKOYA, I. and TOULOUMI, G. (2019). Determining the likely place of HIV acquisition for migrants in Europe combining subject-specific information and biomarkers data. *Stat*. *Methods Med*. *Res*. **28** 1979–1997. MR3977088 https://doi.org/10.1177/0962280217746437

R CORE TEAM (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

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

SHENG, B., LI, C., BAO, L. and LI, R. (2023). Supplement to "Probabilistic HIV recency classification—a logistic regression without labeled individual level training data." https://doi.org/10.1214/22-AOAS1618SUPPA, https://doi.org/10.1214/22-AOAS1618SUPPB

SLAVKOVIC, A. B. (2010). Partial information releases for confidential contingency table entries: Present and future research efforts. *J*. *Priv*. *Confid*. **1**.

SONG, H. and RASKUTTI, G. (2020). PUlasso: High-dimensional variable selection with presence-only data. *J*. *Amer*. *Statist*. *Assoc*. **115** 334–347. MR4078467 https://doi.org/10.1080/01621459.2018.1546587

UNAIDS (2021). AIDSinfo. https://aidsinfo.unaids.org/ (accessed 4 January 2021).

WORLD HEALTH ORGANIZATION (2020). HIV Incidence https://www.who.int/diagnostics_laboratory/links/hiv_incidence_assay/en/ (accessed 29 September 2020).

VAN ENGELEN, J. E. and HOOS, H. H. (2020). A survey on semi-supervised learning. *Mach*. *Learn*. **109** 373–440. MR4070203 https://doi.org/10.1007/s10994-019-05855-6

VAN BUUREN, S. (2018). *Flexible Imputation of Missing Data*. CRC Press, Boca Raton.

WARD, G., HASTIE, T., BARRY, S., ELITH, J. and LEATHWICK, J. R. (2009). Presence-only data and the EM algorithm. *Biometrics* **65** 554–563. MR2751480 https://doi.org/10.1111/j.1541-0420.2008.01116.x

WELTE, A., MCWALTER, T. A. and BÄRNIGHAUSEN, T. (2009). A simplified formula for inferring HIV incidence from cross-sectional surveys using a test for recent infection. *AIDS Res*. *Hum*. *Retrovir*. **25** 125–126.

WESTAT (2018). *Malawi PHIA* (*MPHIA*) 2015-16 *Technical Report*. Columbia Univ., New York.

YANG, X., FIENBERG, S. E. and RINALDO, A. (2012). Differential privacy for protecting multi-dimensional contingency table data: Extensions and applications. *J*. *Priv*. *Confid*. **4**.

ZHOU, Z.-H. (2018). A brief introduction to weakly supervised learning. *Nat*. *Sci*. *Rev*. **5** 44–53.

ZHU, X. and STEPHENS, M. (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann*. *Appl*. *Stat*. **11** 1561–1592. MR3709570 https://doi.org/10.1214/17-AOAS1046

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 130–152
https://doi.org/10.1214/22-AOAS1619

ANALYZING ANIMAL ESCAPE DATA WITH CIRCULAR NONPARAMETRIC MULTIMODAL REGRESSION

BY MARÍA ALONSO-PENA[a] AND ROSA M. CRUJEIRAS[b]

CITMAga, Universidade de Santiago de Compostela, [a]*mariaalonso.pena@usc.es,* [b]*rosa.crujeiras@usc.es*

Analyzing the escape direction of animals subject to covariates is a problem that requires statistical techniques beyond classical regression methods. Apart from the periodicity of the angle of direction, which demands the use of circular statistics, animal escape data usually call for the exploration of the preferred orientations rather than the expected orientation. In this paper we propose the use of a nonparametric method to estimate the conditional local modes of the escape directions of animals from a regression perspective. We present the estimation algorithms and study the asymptotic properties of the estimators as well as its finite sample performance through some simulation experiments. Our proposal is used to model the escape behavior of a group of larval zebrafish escaping from a robot predator. More broadly, the approach presented in this paper can be applied to many existing problems related to animal behavior or other fields.

REFERENCES

ALIYARI GHASSABEH, Y. (2015). A sufficient condition for the convergence of the mean shift algorithm with Gaussian kernel. *J. Multivariate Anal.* **135** 1–10. MR3306422 https://doi.org/10.1016/j.jmva.2014.11.009

ALIYARI GHASSABEH, Y. (2013). On the convergence of the mean shift algorithm in the one-dimensional space. *Pattern Recogn. Lett.* **34** 1423–1427.

ALIYARI GHASSABEH, Y., LINDER, T. and TAKAHARA, G. (2015). On some convergence properties of the subspace constrained mean shift. *Pattern Recognit.* **46** 3140–3147.

ALONSO-PENA, M. and CRUJEIRAS, R. M. (2023a). Supplement to "Analyzing animal escape data with circular nonparametric multimodal regression: Assumptions, results and proofs." https://doi.org/10.1214/22-AOAS1619SUPPA

ALONSO-PENA, M. and CRUJEIRAS, R. M. (2023b). Supplement to "Analyzing animal escape data with circular nonparametric multimodal regression: Modifed modal cross-validation." https://doi.org/10.1214/22-AOAS1619SUPPB

BAI, Z. D., RAO, C. R. and ZHAO, L. C. (1988). Kernel estimators of density function of directional data. *J. Multivariate Anal.* **27** 24–39. MR0971170 https://doi.org/10.1016/0047-259X(88)90113-3

CARD, G. and DICKINSON, M. H. (2008). Visually mediated motor planning in the escape response of Drosophila. *Curr. Biol.* **18** 1300–1307.

CASA, A., CHACÓN, J. E. and MENARDI, G. (2020). Modal clustering asymptotics with applications to bandwidth selection. *Electron. J. Stat.* **14** 835–856. MR4062160 https://doi.org/10.1214/20-EJS1679

CHEN, Y.-C. (2018). Modal regression using kernel density estimation: A review. *Wiley Interdiscip. Rev.: Comput. Stat.* **10** e1431. MR3826098 https://doi.org/10.1002/wics.1431

CHEN, Y. C., GENOVESE, C. R. and WASSERMAN, L. (2014). Generalized mode and ridge estimation. ArXiv E-prints. Available at arXiv:1406.1803.

CHEN, Y.-C., GENOVESE, C. R., TIBSHIRANI, R. J. and WASSERMAN, L. (2016). Nonparametric modal regression. *Ann. Statist.* **44** 489–514. MR3476607 https://doi.org/10.1214/15-AOS1373

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

COMANICIU, D. and MEER, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 603–619.

DI MARZIO, M., PANZERA, A. and TAYLOR, C. C. (2009). Local polynomial regression for circular predictors. *Statist. Probab. Lett.* **79** 2066–2075. MR2571770 https://doi.org/10.1016/j.spl.2009.06.014

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

Key words and phrases. Multimodal regression, circular data, animal escape.

DI MARZIO, M., PANZERA, A. and TAYLOR, C. C. (2013). Non-parametric regression for circular responses. *Scand. J. Stat.* **40** 238–255. MR3066413 https://doi.org/10.1111/j.1467-9469.2012.00809.x

DI MARZIO, M., FENSORE, S., PANZERA, A. and TAYLOR, C. C. (2016). A note on nonparametric estimation of circular conditional densities. *J. Stat. Comput. Simul.* **86** 2573–2582. MR3511014 https://doi.org/10.1080/00949655.2016.1146279

DOMENICI, P., BLAGBURN, J. M. and BACON, J. P. (2011a). Animal escapology I: Theoretical issues and emerging trends in escape trajectories. *J. Exp. Biol.* **214** 2463–2473.

DOMENICI, P., BLAGBURN, J. M. and BACON, J. P. (2011b). Animal escapology II: Escape trajectory case studies. *J. Exp. Biol.* **214** 2474–2494.

EDELSBRUNNER, H., KIRKPATRICK, D. G. and SEIDEL, R. (1983). On the shape of a set of points in the plane. *IEEE Trans. Inf. Theory* **29** 551–559. MR0713690 https://doi.org/10.1109/TIT.1983.1056714

EDELSBRUNNER, H. and MUCKE, E. P. (1984). Three-dimensional alpha shapes. *ACM Trans. Graph.* **13** 43–72.

EINBECK, J. and TUTZ, G. (2006). Modelling beyond regression functions: An application of multimodal regression to speed-flow data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **55** 461–475. MR2242274 https://doi.org/10.1111/j.1467-9876.2006.00547.x

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications. Monographs on Statistics and Applied Probability* **66**. CRC Press, London. MR1383587

FISHER, N. I. (1989). Smoothing a sample of circular data. *J. Struct. Geol.* **11** 775–778.

FISHER, N. I. (1993b). *Statistical Analysis of Circular Data*. Cambridge Univ. Press, Cambridge. MR1251957 https://doi.org/10.1017/CBO9780511564345

FISHER, N. I. and LEE, A. J. (1992). Regression models for an angular response. *Biometrics* **48** 665–677. MR1187598 https://doi.org/10.2307/2532334

FUKUNAGA, K. and HOSTETLER, L. D. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory* **21** 32–40. MR0388638 https://doi.org/10.1109/tit.1975.1055330

GARCÍA-PORTUGUÉS, E., CRUJEIRAS, R. M. and GONZÁLEZ-MANTEIGA, W. (2013). Kernel density estimation for directional-linear data. *J. Multivariate Anal.* **121** 152–175. MR3090475 https://doi.org/10.1016/j.jmva.2013.06.009

JAMMALAMADAKA, S. R. and SENGUPTA, A. (2001). *Topics in Circular Statistics. Series on Multivariate Analysis* **5**. World Scientific Co., Inc., River Edge, NJ. MR1836122 https://doi.org/10.1142/9789812779267

KIM, S. and SENGUPTA, A. (2017). Multivariate-multiple circular regression. *J. Stat. Comput. Simul.* **87** 1277–1291. MR3621948 https://doi.org/10.1080/00949655.2016.1261292

KOBAYASHI, T. and OTSU, N. (2010). von Mises-Fisher mean shift for clustering on a hypersphere. In *Proceedings of the* 20*th International Conference on Pattern Recognition* 2130–2133. IEEE, New York.

LEY, C. and VERDEBOUT, T. (2017). *Modern Directional Statistics. Chapman & Hall/CRC Interdisciplinary Statistics Series*. CRC Press, Boca Raton, FL. MR3752655

LI, X., HU, Z. and WU, F. (2007). A note on the convergence of the mean shift. *Pattern Recognit.* **40** 1756–1762.

MARCHETTI, G. and SCAPINI, F. (2003). Use of multiple regression models in the study of sandhopper orientation under natural conditions. *Estuar. Coast. Shelf Sci.* **58** 207–215.

MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics. Wiley Series in Probability and Statistics*. Wiley, Chichester. MR1828667

NAIR, A., CHANGSING, K., STEWART, W. J. and MCHENRY, M. J. (2017a). Data from: Fish prey change strategy with the direction of a threat. https://doi.org/10.5061/dryad.47mq9.

NAIR, A., CHANGSING, K., STEWART, W. J. and MCHENRY, M. J. (2017b). Fish prey change strategy with the direction of a threat. *Proc. Biol. Sci.* **284**. https://doi.org/10.1098/rspb.2017.0393

OBA, S., KATO, K. and ISHII, S. (2005). Multi-scale clustering for gene expression profiling data. In 5*th IEEE Symposium on Bioinformatics and Bioengineering* (*BIBE*'05) 210–217. IEEE, New York.

OBLESER, P., HART, V., MALKEMPER, E. P., BEGALL, S., HOLÁ, M., PAINTER, M. S., CERVENÝ, J. and BURDA, H. (2016). Compass-controlled escape behavior in Roe deer. *Behav. Ecol. Sociobiol.* **70** 1345–1355.

OLIVEIRA, M., CRUJEIRAS, R. M. and RODRÍGUEZ–CASAL, A. (2014). NPCirc: An R package for nonparametric circular methods. *J. Stat. Softw.* **61**.

PEWSEY, A., NEUHÄUSER, M. and RUXTON, G. D. (2013). *Circular Statistics in R*. Oxford Univ. Press, Oxford. MR3156170

SATO, N., SHIDARA, H. and OGAWA, H. (2019). Trade-off between motor performance and behavioural flexibility in the action selection of cricket escape behaviour. *Sci. Rep.* **9** 18112. https://doi.org/10.1038/s41598-019-54555-7

SCAPINI, F., ALOIA, A., BOUSLAMA, M., CHELAZZI, L., COLOMBINI, I., ELGTARI, M., FALLACI, M. and MARCHETTI, G. M. (2002). Multiple regression analysis of the sources of variation in orientation of two sympatric sandhoppers, talitrus saltator and talorchestia brito, from an exposed Mediterranean beach. *Behav. Ecol.* **52** 403–414.

SCOTT, D. W. (1992). *Multivariate Density Estimation*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1191168 https://doi.org/10.1002/9780470316849

WIKIMEDIA COMMONS (2014). Larval zebrafish. [Online; accessed September 29, 2021].

WILLIAMSON, B., FRASER, S., WILLIAMSON, L., NIKORA, V. and SCOTT, B. (2019). Predictable changes in fish school characteristics due to a tidal turbine support structure. *Renew. Energy* **141** 1092–1102.

ZELNIK-MANOR, L. and PERONA, P. (2004). Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems* (*NIPS*).

ZHANG, Y. and CHEN, Y.-C. (2021a). Kernel smoothing, mean shift, and their learning theory with directional data. *J. Mach. Learn. Res.* **22** Paper No. 154. MR4318510

ZHANG, Y. and CHEN, Y. C. (2021b). Mode and ridge estimation in euclidean and directional product spaces: A mean shift approach. ArXiv E-prints. Available at arXiv:2110.08505.

ZHOU, H. and HUANG, X. (2016). Nonparametric modal regression in the presence of measurement error. *Electron. J. Stat.* **10** 3579–3620. MR3575565 https://doi.org/10.1214/16-EJS1210

ZHOU, H. and HUANG, X. (2019). Bandwidth selection for nonparametric modal regression. *Comm. Statist. Simulation Comput.* **48** 968–984. MR3957561 https://doi.org/10.1080/03610918.2017.1402044

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 153–177
https://doi.org/10.1214/22-AOAS1620

SEQUENTIAL SAMPLING IN PROSPECTIVE OBSERVATIONAL STUDIES

BY MARY M. RYAN[a] AND DANIEL L. GILLEN[b]

Department of Statistics, University of California, Irvine, [a]*marymr@uci.edu,* [b]*dgillen@uci.edu*

Many factors must be taken into account when designing an observational study. Unlike controlled studies, observational studies cannot mitigate the effects of confounding through randomization, and such factors should be incorporated into both the study analysis and the study design. Unfortunately, there is often little data available on most of these factors at the design stage, rendering it infeasible to reliably postulate the impact of these factors on the treatment effect estimate and precision of such an effect. In this work we demonstrate how failure to account for adjustment covariates in the design stage of observational studies utilizing group sequential designs has deleterious effects on the study's observed power. We do this by constructing a hypothetical study of Alzheimer's disease biomarkers on cognition, using data from the Alzheimer's Disease Neuroimaging Initiative. We then outline a constrained boundaries procedure that uses data collected at interim analyses to better estimate variance and correct stopping boundaries to maintain power and type I error, while allowing for early study termination.

REFERENCES

ARMITAGE, P., MCPHERSON, C. K. and ROWE, B. C. (1969). Repeated significance tests on accumulating data. *J. Roy. Statist. Soc. Ser. A* **132** 235–244. MR0250405 https://doi.org/10.2307/2343787

BURINGTON, B. E. and EMERSON, S. S. (2003). Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics* **59** 770–777. MR2025101 https://doi.org/10.1111/j.0006-341X.2003.00090.x

CORDER, E. H., SAUNDERS, A. M., STRITTMATTER, W. J., SCHMECHEL, D. E., GASKELL, P. C., SMALL, G. W., ROSES, A. D., HAINES, J. L. and PERICAK-VANCE, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261** 921–923.

EMERSON, S. S. and FLEMING, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77** 875–892. MR1086697 https://doi.org/10.1093/biomet/77.4.875

EMERSON, S. S., KITTELSON, J. M. and GILLEN, D. L. (2007). Frequentist evaluation of group sequential clinical trial designs. *Stat. Med.* **26** 5047–5080. MR2412457 https://doi.org/10.1002/sim.2901

EMERSON, S. S., GILLEN, D. L., KITTELSON, J. M., EMERSON, S. C. and LEVIN, G. P. RCTdesign.org: Methods and Software for Clinical Trials.

GILLEN, D. L. (2003). *The Use of Weighted Logrank Statistics in Group Sequential Testing and Nonproportional Hazards*. University of Washington.

HANEUSE, S., SCHILDCROUT, J. and GILLEN, D. (2012). A two-stage strategy to accommodate general patterns of confounding in the design of observational studies. *Biostatistics* **13** 274–288.

HARDY, J. A. and HIGGINS, G. A. (1992). Alzheimer's disease: The amyloid cascade hypothesis. *Science* **256** 184.

HAYBITTLE, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *The British Journal of Radiology* **44** 793–797.

HUGHES, M. D. and POCOCK, S. J. (1988). Stopping rules and estimation problems in clinical trials. *Stat. Med.* **7** 1231–1242.

JENNISON, C. and TURNBULL, B. W. (1984). Repeated confidence intervals for group sequential clinical trials. *Control. Clin. Trials* **5** 33–45.

JENNISON, C. and TURNBULL, B. W. (1997). Group-sequential analysis incorporating covariate information. *J. Amer. Statist. Assoc.* **92** 1330–1341. MR1615245 https://doi.org/10.2307/2965403

KARCH, C. M. and GOATE, A. M. (2015). Alzheimer's disease risk genes and mechanisms of disease pathogenesis. *Biol. Psychiatry* **77** 43–51. https://doi.org/10.1016/j.biopsych.2014.05.006

Key words and phrases. Group sequential, adaptive design, observational study design, power, sample size, confounding.

KITTELSON, J. M. and EMERSON, S. S. (1999). A unifying family of group sequential test designs. *Biometrics* **55** 874–882. MR1720716 https://doi.org/10.1111/j.0006-341X.1999.00874.x

MOHS, R. C., KNOPMAN, D., PETERSEN, R. C., FERRIS, S. H., ERNESTO, C., GRUNDMAN, M., SANO, M., BIELIAUSKAS, L., GELDMACHER, D. et al. (1997). Development of cognitive instruments for use in clinical trials of antidementia drugs: Additions to the Alzheimer's disease assessment scale that broaden its scope. The Alzheimer's disease cooperative study. *Alzheimer Disease and Associated Disorders* **11** S13–21.

MURRAY, S. and TSIATIS, A. A. (1999). Sequential methods for comparing years of life saved in the two-sample censored data problem. *Biometrics* **55** 1085–1092.

O'BRIEN, P. C. and FLEMING, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35** 549–556.

PETO, R., PIKE, M. C., ARMITAGE, P., BRESLOW, N. E., COX, D. R., HOWARD, S. V., MANTEL, N., MCPHERSON, K., PETO, J. et al. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br. J. Cancer* **34** 585–612. https://doi.org/10.1038/bjc.1976.220

POCOCK, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64** 191–199.

READY, R., OTT, B. and GRACE, J. (2004). Validity of informant reports about AD and MCI patients' memory. *Alzheimer Disease & Associated Disorders* **18** 11–16.

ROYCHOUDHURY, S., ANDERSON, K. M., YE, J. and MUKHOPADHYAY, P. (2021). Robust design and analysis of clinical trials with nonproportional hazards: A straw man guidance from a cross-pharma working group. *Stat. Biopharm. Res.* **0** 1–15.

RYAN, M. M. and GILLEN, D. L. (2023). Supplement to "Sequential sampling in prospective observational studies." https://doi.org/10.1214/22-AOAS1620SUPPA, https://doi.org/10.1214/22-AOAS1620SUPPB

TARIMA, S., HE, P., WANG, T. and SZABO, A. (2016). An interim sample size recalculation for observational studies. *Observational Studies* **2** 65–85.

TSIATIS, A. A., ROSNER, G. L. and MEHTA, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40** 797–803. MR0775387 https://doi.org/10.2307/2530924

WHITEHEAD, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **73** 573–581. MR0897848 https://doi.org/10.1093/biomet/73.3.573

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 178–198
https://doi.org/10.1214/22-AOAS1621

CONEX–CONNECT: LEARNING PATTERNS IN EXTREMAL BRAIN CONNECTIVITY FROM MULTICHANNEL EEG DATA

BY MATHEUS B. GUERRERO[a], RAPHAËL HUSER[b] AND HERNANDO OMBAO[c]

Statistics Program, CEMSE Division, King Abdullah University of Science and Technology (KAUST),
[a]*matheus.bartologuerrero@kaust.edu.sa,* [b]*raphael.huser@kaust.edu.sa,* [c]*hernando.ombao@kaust.edu.sa*

Epilepsy is a chronic neurological disorder; it affects more than 50 million people globally. An epileptic seizure acts like a temporary shock to the neuronal system, disrupting normal electrical activity in the brain. Epilepsy is frequently diagnosed with electroencephalograms (EEGs). Current methods study only the time-varying spectra and coherence but do not directly model changes in extreme behavior, neglecting the fact that neuronal oscillations exhibit non-Gaussian heavy-tailed probability distributions. To overcome this limitation, we propose a new approach to characterize brain connectivity based on the joint tail (i.e., extreme) behavior of the EEGs. Our proposed method, the conditional extremal dependence for brain connectivity (Conex–Connect), is a pioneering approach that links the association between extreme values of higher oscillations at a reference channel with the other brain network channels. Using the Conex–Connect method, we discover changes in the extremal dependence driven by the activity at the foci of the epileptic seizure. Our model-based approach reveals that, preseizure, the dependence is notably stable for all channels when conditioning on extreme values of the focal seizure area. By contrast, the dependence between channels is weaker during the seizure, and dependence patterns are more "chaotic." Using the Conex–Connect method, we identified the high-frequency oscillations as the most relevant features, explaining the conditional extremal dependence of brain connectivity.

REFERENCES

ACHARYA, U. R., VINITHA SREE, S., SWAPNA, G., MARTIS, R. J. and SURI, J. S. (2013). Automated EEG analysis of epilepsy: A review. *Knowl.-Based Syst*. **45** 147–165.

BOWYER, S. M. (2016). Coherence a measure of the brain networks: Past and present. *Neuropsychiatr. Electrophysiol*. **2** 1.

COHEN, M. X. (2014). *Analyzing Neural Time Series Data*. MIT Press, USA.

COLES, S. G., HEFFERNAN, J. and TAWN, J. A. (1999). Dependence measures for extreme value analyses. *Extremes* **2** 339–365.

DAS, B. and RESNICK, S. I. (2011). Conditioning on an extreme component: Model consistency with regular variation on cones. *Bernoulli* **17** 226–252. MR2797990 https://doi.org/10.3150/10-BEJ271

DAVIS, R. A. and MIKOSCH, T. (2009). The extremogram: A correlogram for extreme events. *Bernoulli* **15** 977–1009. MR2597580 https://doi.org/10.3150/09-BEJ213

DAVIS, R. A., MIKOSCH, T. and CRIBBEN, I. (2012). Towards estimating extremal serial dependence via the bootstrapped extremogram. *J. Econometrics* **170** 142–152. MR2955945 https://doi.org/10.1016/j.jeconom.2012.04.003

DAVISON, A. C. (2003). *Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **11**. Cambridge Univ. Press, Cambridge. MR1998913 https://doi.org/10.1017/CBO9780511815850

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

DAVISON, A. C. and HUSER, R. (2015). Statistics of extremes. *Annu. Rev. Stat. Appl*. **2** 203–235.

EWANS, K. and JONATHAN, P. (2014). Evaluating environmental joint extremes for the offshore industry using the conditional extremes model. *J. Mar. Syst*. **130** 124–130.

Key words and phrases. Epilepsy, extreme-value theory, conditional extremes, penalized likelihood, nonstationary time series.

FIECAS, M. and OMBAO, H. (2016). Modeling the evolution of dynamic brain processes during an associative learning experiment. *J. Amer. Statist. Assoc.* **111** 1440–1453. MR3601700 https://doi.org/10.1080/01621459.2016.1165683

GAO, X., SHEN, W., SHAHBABA, B., FORTIN, N. J. and OMBAO, H. (2020). Evolutionary state-space model and its application to time-frequency analysis of local field potentials. *Statist. Sinica* **30** 1561–1582. MR4257545 https://doi.org/10.5705/ss.202017.0420

GUERRERO, M. B., HUSER, R. and OMBAO, H. (2023a). Supplement to "Conex–Connect: Learning patterns in extremal brain connectivity from multichannel EEG data." https://doi.org/10.1214/22-AOAS1621SUPPA

GUERRERO, M. B., HUSER, R. and OMBAO, H. (2023b). R code for "Conex–Connect: Learning patterns in extremal brain connectivity from multichannel EEG data." https://doi.org/10.1214/22-AOAS1621SUPPB

HEFFERNAN, J. E. and RESNICK, S. I. (2007). Limit laws for random vectors with an extreme component. *Ann. Appl. Probab.* **17** 537–571. MR2308335 https://doi.org/10.1214/105051606000000835

HEFFERNAN, J. E. and TAWN, J. A. (2004). A conditional approach for multivariate extreme values. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 497–546. MR2088289 https://doi.org/10.1111/j.1467-9868.2004.02050.x

HILAL, S., POON, S. H. and TAWN, J. (2011). Hedging the black swan: Conditional heteroskedasticity and tail dependence in S&P500 and VIX. *J. Bank. Financ.* **35** 2374–2387.

HYNDMAN, R. J. and ATHANASOPOULOS, G. (2019). *Forecasting: Principles and Practice*, 3rd ed. OTexts, Australia. Available at https://OTexts.com/fpp3. Accessed on 08/20/2020.

KEEF, C., PAPASTATHOPOULOS, I. and TAWN, J. A. (2013). Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model. *J. Multivariate Anal.* **115** 396–404. MR3004566 https://doi.org/10.1016/j.jmva.2012.10.012

KRAFTY, R. T. and COLLINGE, W. O. (2013). Penalized multivariate Whittle likelihood for power spectrum estimation. *Biometrika* **100** 447–458. MR3068445 https://doi.org/10.1093/biomet/ass088

KRAFTY, R. T., HALL, M. and GUO, W. (2011). Functional mixed effects spectral analysis. *Biometrika* **98** 583–598. MR2836408 https://doi.org/10.1093/biomet/asr032

KRAFTY, R. T., ROSEN, O., STOFFER, D. S., BUYSSE, D. J. and HALL, M. H. (2017). Conditional spectral analysis of replicated multiple time series with application to nocturnal physiology. *J. Amer. Statist. Assoc.* **112** 1405–1416. MR3750864 https://doi.org/10.1080/01621459.2017.1281811

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

LEDFORD, A. W. and TAWN, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika* **83** 169–187. MR1399163 https://doi.org/10.1093/biomet/83.1.169

LUGRIN, T., DAVISON, A. C. and TAWN, J. A. (2019). Penultimate analysis of the conditional multivariate extremes tail model. Available at arXiv:1902.06972.

MEDVEDEV, A. V., MURRO, A. M. and MEADOR, K. J. (2011). Abnormal interictal gamma activity may manifest a seizure onset zone in temporal lobe epilepsy. *Int. J. Neural Syst.* **21** 103–114.

NUNEZ, P. L. and SRINIVASAN, R. (2007). Electroencephalogram. *Scholarpedia* **2** 1348.

OMBAO, H. and PINTO, M. (2021). Spectral dependence. Available at arXiv:2103.17240.

OMBAO, H. and VAN BELLEGEM, S. (2008). Evolutionary coherence of nonstationary signals. *IEEE Trans. Signal Process.* **56** 2259–2266. MR2516630 https://doi.org/10.1109/TSP.2007.914341

OMBAO, H., VON SACHS, R. and GUO, W. (2005). SLEX analysis of multivariate nonstationary time series. *J. Amer. Statist. Assoc.* **100** 519–531. MR2160556 https://doi.org/10.1198/016214504000001448

OMBAO, H. C., RAZ, J. A., VON SACHS, R. and MALOW, B. A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *J. Amer. Statist. Assoc.* **96** 543–560. MR1946424 https://doi.org/10.1198/016214501753168244

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

OMBAO, H., FIECAS, M., TING, C.-M. and LOW, Y. F. (2018). Statistical models for brain signals with properties that evolve across trials. *NeuroImage* **180** 609–618. https://doi.org/10.1016/j.neuroimage.2017.11.061

ROSS, E., SAM, S., RANDELL, D., FELD, G. and JONATHAN, P. (2018). Estimating surge in extreme North Sea storms. *Ocean Engineering* **154** 430–444.

SCHEFFLER, A. W., DICKINSON, A., DISTEFANO, C., JESTE, S. and ŞENTÜRK, D. (2020). Covariate-adjusted hybrid principal components analysis. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems* (M. J. Lesot, S. Vieira, M. Z. Reformat, J. P. Carvalho, A. Wilbik, B. Bouchon-Meunier and R. R. Yager, eds.) Springer, Switzerland.

SCHRÖDER, A. L. and OMBAO, H. (2019). FreSpeD: Frequency-specific change-point detection in epileptic seizure multi-channel EEG data. *J. Amer. Statist. Assoc.* **114** 115–128. MR3941242 https://doi.org/10.1080/01621459.2018.1476238

TAWN, J., SHOOTER, R., TOWE, R. and LAMB, R. (2018). Modelling spatial extreme events with environmental applications. *Spat. Stat.* **28** 39–58. MR3887155 https://doi.org/10.1016/j.spasta.2018.04.007

WADSWORTH, J. L. and TAWN, J. (2019). Higher-dimensional spatial extremes via single-site conditioning. Available at arXiv:1912.06560.

WHO (2019). Epilepsy: A public health imperative—summary. Available at https://www.who.int/mental_health/neurology/epilepsy/report_2019/en/. Accessed on 05/20/2020.

ZHANG, Z. (2008). Quotient correlation: A simple based alternative to Pearson's correlation. *Ann. Statist.* **36** 1007–1030. MR2396823 https://doi.org/10.1214/009053607000000866

ZHANG, Z. (2021). On studying extreme values and systematic risks with nonlinear time series models and tail dependence measures. *Stat. Theory Relat. Fields* **5** 1–25. MR4232528 https://doi.org/10.1080/24754269.2020.1856590

ZHANG, Z., ZHANG, C. and CUI, Q. (2017). Random threshold driven tail dependence measures with application to precipitation data analysis. *Statist. Sinica* **27** 685–709. MR3674692

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 199–224
https://doi.org/10.1214/22-AOAS1623

BAYESIAN CLASSIFICATION, ANOMALY DETECTION, AND SURVIVAL ANALYSIS USING NETWORK INPUTS WITH APPLICATION TO THE MICROBIOME

BY NATHANIEL JOSEPHS[1,a], LIZHEN LIN[2,b], STEVEN ROSENBERG[3,c] AND ERIC D. KOLACZYK[3,d]

[1]*Department of Biostatistics, Yale University,* [a]*nathaniel.josephs@yale.edu*
[2]*Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame,* [b]*lizhen.lin@nd.edu*
[3]*Department of Mathematics and Statistics, Boston University,* [c]*sr@bu.edu,* [d]*kolaczyk@bu.edu*

While the study of a single network is well established, technological advances now allow for the collection of multiple networks with relative ease. Increasingly, anywhere from several to thousands of networks can be created from brain imaging, gene coexpression data, or microbiome measurements. And these networks, in turn, are being looked to as potentially powerful features to be used in modeling. However, with networks being non-Euclidean in nature, how best to incorporate them into standard modeling tasks is not obvious. In this paper we propose a Bayesian modeling framework that provides a unified approach to binary classification, anomaly detection, and survival analysis with network inputs. We encode the networks in the kernel of a Gaussian process prior via their pairwise differences, and we discuss several choices of provably positive definite kernel that can be plugged into our models. Although our methods are widely applicable, we are motivated here, in particular, by microbiome research (where network analysis is emerging as the standard approach for capturing the interconnectedness of microbial taxa across both time and space) and its potential for reducing preterm delivery and improving personalization of prenatal care.

REFERENCES

ARROYO RELIÓN, J. D., KESSLER, D., LEVINA, E. and TAYLOR, S. F. (2019). Network classification with applications to brain connectomics. *Ann. Appl. Stat.* **13** 1648–1677. MR4019153 https://doi.org/10.1214/19-AOAS1252

ARROYO, J., ATHREYA, A., CAPE, J., CHEN, G., PRIEBE, C. E. and VOGELSTEIN, J. T. (2021). Inference for multiple heterogeneous networks with a common invariant subspace. *J. Mach. Learn. Res.* **22** Paper No. 142, 49. MR4318498

BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634 https://doi.org/10.1126/science.286.5439.509

BOGART, E., CRESWELL, R. and GERBER, G. K. (2019). MITRE: Inferring features from microbiota time-series data linked to host status. *Genome Biol.* **20** 1–15.

BORGWARDT, K. M., ONG, C. S., SCHÖNAUER, S., VISHWANATHAN, S. V. N., SMOLA, A. J. and KRIEGEL, H.-P. (2005). Protein function prediction via graph kernels. *Bioinformatics* **21** i47–i56. https://doi.org/10.1093/bioinformatics/bti1007

CORNEA, E., ZHU, H., KIM, P. and IBRAHIM, J. G. (2017). Regression models on Riemannian symmetric spaces. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 463–482. MR3611755 https://doi.org/10.1111/rssb.12169

DAI, X. and MÜLLER, H.-G. (2018). Principal component analysis for functional data on Riemannian manifolds and spheres. *Ann. Statist.* **46** 3334–3361. MR3852654 https://doi.org/10.1214/17-AOS1660

DE IORIO, M., JOHNSON, W. O., MÜLLER, P. and ROSNER, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics* **65** 762–771. MR2649849 https://doi.org/10.1111/j.1541-0420.2008.01166.x

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

Key words and phrases. Multiple networks, Gaussian process, graph kernel, Frobenius distance, spectral distance, random walk kernel, microbiome networks, preterm delivery.

DONNAT, C. and HOLMES, S. (2018). Tracking network dynamics: A survey using graph distances. *Ann. Appl. Stat.* **12** 971–1012. MR3834292 https://doi.org/10.1214/18-AOAS1176

DURANTE, D., DUNSON, D. B. and VOGELSTEIN, J. T. (2017). Nonparametric Bayes modeling of populations of networks. *J. Amer. Statist. Assoc.* **112** 1516–1530. MR3750873 https://doi.org/10.1080/01621459.2016.1219260

FERNÁNDEZ, T., RIVERA, N. and TEH, Y. W. (2016). Gaussian processes for survival analysis. In *Advances in Neural Information Processing Systems* 5021–5029.

FERNÁNDEZ, T. and TEH, Y. W. (2016). Posterior consistency for a non-parametric survival model under a Gaussian process prior. Preprint. Available at arXiv:1611.02335.

FRIEDMAN, J. and ALM, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8** e1002687. https://doi.org/10.1371/journal.pcbi.1002687

GÄRTNER, T., DRIESSENS, K. and RAMON, J. (2003). Graph kernels and Gaussian processes for relational reinforcement learning. In *International Conference on Inductive Logic Programming* 146–163. Springer.

GHOSAL, S. and ROY, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. *Ann. Statist.* **34** 2413–2429. MR2291505 https://doi.org/10.1214/009053606000000795

GINESTET, C. E., LI, J., BALACHANDRAN, P., ROSENBERG, S. and KOLACZYK, E. D. (2017). Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.* **11** 725–750. MR3693544 https://doi.org/10.1214/16-AOAS1015

GOLLINI, I. and MURPHY, T. B. (2016). Joint modeling of multiple network views. *J. Comput. Graph. Statist.* **25** 246–265. MR3474046 https://doi.org/10.1080/10618600.2014.978006

HSU, C.-W., CHANG, C.-C., LIN, C.-J. et al. (2003). A practical guide to support vector classification.

JAIN, B. J. (2016). On the geometry of graph spaces. *Discrete Appl. Math.* **214** 126–144. MR3546275 https://doi.org/10.1016/j.dam.2016.06.027

JAYASUMANA, S., HARTLEY, R., SALZMANN, M., LI, H. and HARANDI, M. (2013). Kernel methods on the Riemannian manifold of symmetric positive definite matrices. In *Computer Vision and Pattern Recognition* (*CVPR*), 2013 *IEEE Conference on* 73–80. IEEE Press, New York.

JOSEPHS, N., LIN, L., ROSENBERG, S. and KOLACZYK, E. D. (2023). Supplement to "Bayesian classification, anomaly detection, and survival analysis using network inputs with application to the microbiome." https://doi.org/10.1214/22-AOAS1623SUPPA, https://doi.org/10.1214/22-AOAS1623SUPPB

KASHIMA, H. and INOKUCHI, A. (2002). Kernels for graph classification. In *ICDM Workshop on Active Mining* **2002**.

KEMMLER, M., RODNER, E., WACKER, E.-S. and DENZLER, J. (2013). One-class classification with Gaussian processes. *Pattern Recognit.* **46** 3507–3518.

KHAN, S. S. and MADDEN, M. G. (2009). A survey of recent trends in one class classification. In *Irish Conference on Artificial Intelligence and Cognitive Science* 188–197. Springer, Berlin.

KIM, M. and PAVLOVIC, V. (2018). Variational inference for Gaussian process models for survival analysis. In *UAI* 435–445.

KOLACZYK, E. D. and CSÁRDI, G. (2014). *Statistical Analysis of Network Data with R. Use R!* Springer, New York. MR3288852 https://doi.org/10.1007/978-1-4939-0983-4

KOLACZYK, E. D., LIN, L., ROSENBERG, S., WALTERS, J. and XU, J. (2020). Averages of unlabeled networks: Geometric characterization and asymptotic behavior. *Ann. Statist.* **48** 514–538. MR4065172 https://doi.org/10.1214/19-AOS1820

KONDOR, R. I. and LAFFERTY, J. (2002). Diffusion kernels on graphs and other discrete structures. In *Proceedings of the* 19*th International Conference on Machine Learning* **2002** 315–322.

KRIEGE, N. M., JOHANSSON, F. D. and MORRIS, C. (2020). A survey on graph kernels. *Appl. Netw. Sci.* **5** 1–42.

KUNEGIS, J., SCHMIDT, S., LOMMATZSCH, A., LERNER, J., DE LUCA, E. W. and ALBAYRAK, S. (2010). Spectral analysis of signed graphs for clustering, prediction and visualization. In *Proceedings of the* 2010 *SIAM International Conference on Data Mining* 559–570. SIAM, Philadelphia.

LAYEGHIFARD, M., HWANG, D. M. and GUTTMAN, D. S. (2017). Disentangling interactions in the microbiome: A network perspective. *Trends Microbiol.* **25** 217–228. https://doi.org/10.1016/j.tim.2016.11.008

LUNAGÓMEZ, S., OLHEDE, S. C. and WOLFE, P. J. (2021). Modeling network populations via graph distances. *J. Amer. Statist. Assoc.* **116** 2023–2040. MR4353730 https://doi.org/10.1080/01621459.2020.1763803

MCNEISH, D. (2016). On using Bayesian methods to address small sample problems. *Struct. Equ. Model.* **23** 750–773. MR3548112 https://doi.org/10.1080/10705511.2016.1186549

MUKHERJEE, S. S., SARKAR, P. and LIN, L. (2017). On clustering network-valued data. In *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) **30** 7071–7081.

MURRAY, I. and ADAMS, R. P. (2010). Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems* 1732–1740.

MURRAY, I., ADAMS, R. P. and MACKAY, D. J. (2010). Elliptical slice sampling.
MYGDALIS, V., IOSIFIDIS, A., TEFAS, A. and PITAS, I. (2016). Graph embedded one-class classifiers for media data classification. *Pattern Recognit*. **60** 585–595.
NIKOLENTZOS, G., SIGLIDIS, I. and VAZIRGIANNIS, M. (2021). Graph kernels: A survey. *J. Artificial Intelligence Res*. **72** 943–1027. MR4366693
OKSANEN, J. (2013). Vegan: Ecological diversity. *R Project* **368**.
PEDARSANI, P. and GROSSGLAUSER, M. (2011). On the privacy of anonymized networks. In *Proceedings of the* 17*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1235–1243.
RAHIMI, A. and RECHT, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems* 1177–1184.
RALAIVOLA, L., SWAMIDASS, S. J., SAIGO, H. and BALDI, P. (2005). Graph kernels for chemical informatics. *Neural Netw*. **18** 1093–1110.
RAMON, E., BELANCHE-MUÑOZ, L., MOLIST, F., QUINTANILLA, R., PEREZ-ENCISO, M. and RAMAYO-CALDAS, Y. (2021). KernInt: A kernel framework for integrating supervised and unsupervised analyses in spatio-temporal metagenomic datasets. *Front*. *Microbiol*. **12** 60.
RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435
RUDD, J. M. (2018). Application of support vector machine modeling and graph theory metrics for disease classification. *Model Assist*. *Stat*. *Appl*. **13** 341–349.
SALTER-TOWNSHEND, M. and MCCORMICK, T. H. (2017). Latent space models for multiview network data. *Ann*. *Appl*. *Stat*. **11** 1217–1244. MR3709558 https://doi.org/10.1214/16-AOAS955
TANG, R., KETCHA, M., BADEA, A., CALABRESE, E. D., MARGULIES, D. S., VOGELSTEIN, J. T., PRIEBE, C. E. and SUSSMAN, D. L. (2018). Connectome smoothing via low-rank approximations. *IEEE Trans*. *Med*. *Imag*. **38** 1446–1456.
VISHWANATHAN, S. V. N., SCHRAUDOLPH, N. N., KONDOR, R. and BORGWARDT, K. M. (2010). Graph kernels. *J*. *Mach*. *Learn*. *Res*. **11** 1201–1242. MR2645450 https://doi.org/10.1093/chemse/bjq147
WATTS, D. J. and STROGATZ, S. H. (1998). Collective dynamics of ‘small-world’networks. *Nature* **393** 440–442.
WILLS, P. and MEYER, F. G. (2020). Metrics for graph comparison: A practitioner’s guide. *PLoS ONE* **15** e0228728. https://doi.org/10.1371/journal.pone.0228728
ZHANG, W., OTA, T., SHRIDHAR, V., CHIEN, J., WU, B. and KUANG, R. (2013). Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS Comput*. *Biol*. **9** e1002975.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 225–239
https://doi.org/10.1214/22-AOAS1624

REGULARIZED FINGERPRINTING IN DETECTION AND ATTRIBUTION OF CLIMATE CHANGE WITH WEIGHT MATRIX OPTIMIZING THE EFFICIENCY IN SCALING FACTOR ESTIMATION

BY YAN LI[1,a], KUN CHEN[1,b], JUN YAN[1,c] AND XUEBIN ZHANG[2,d]

[1]*Department of Statistics, University of Connecticut,* [a]*yan.4.li@uconn.edu,* [b]*kun.chen@uconn.edu,* [c]*jun.yan@uconn.edu*
[2]*Division of Climate Research, Environmental and Climate Change Canada,* [d]*xuebin.zhang@canada.ca*

Detection and attribution analyses play a central role in establishing the causal effect of human activities on global warming. The most commonly used method in such analyses, optimal fingerprinting, is a multiple regression where each covariate has a measurement error whose covariance matrix is the same as that of the regression error up to a known scale. Inferences about the regression coefficients are critical not only for making statements about detection and attribution but also for quantifying the uncertainty in important outcomes derived from detection and attribution analyses. When there are no errors-in-variables (EIV), the optimal weight matrix in estimating the regression coefficients is the precision matrix of the regression error. This matrix, however, is never known and has to be estimated from climate model simulations with regularization. The consequence is that the optimal fingerprinting method is not optimal, as believed in practice. We construct a weight matrix by inverting a nonlinear shrinkage estimate of the error covariance matrix that minimizes loss functions directly targeting the uncertainty of the resulting regression coefficient estimator. The resulting estimator of the regression coefficients is asymptotically optimal as the sample size of the climate model simulations and the matrix dimension go to infinity together with a limiting ratio. When EIVs are present, the estimator of the regression coefficients, based on the proposed weight matrix, is asymptotically more efficient than that based on the inverse of the existing linear shrinkage estimator of the error covariance matrix. The performance of the method is confirmed in finite sample simulation studies mimicking realistic situations in terms of the length of the confidence intervals and empirical coverage rates for the regression coefficients. In an application to detection and attribution analyses of the mean temperature at different spatial scales, the method yielded shorter confidence intervals which are important for such analyses in practice.

REFERENCES

ALLEN, M. R. and STOTT, P. A. (2003). Estimating signal amplitudes in optimal fingerprinting, Part I: Theory. *Clim. Dyn.* **21** 477–491.

ALLEN, M. R. and TETT, S. F. B. (1999). Checking for model consistency in optimal fingerprinting. *Clim. Dyn.* **15** 419–434.

BATEMAN, H. (1954). *Tables of Integral Transforms* **II**. McGraw-Hill, New York.

BINDOFF, N. L., STOTT, P. A., ACHUTARAO, K. M., ALLEN, M. R., GILLETT, N., GUTZLER, D., HANSINGO, K., HEGERL, G., HU, Y. et al. (2013). Detection and attribution of climate change: From global to regional. In *Climate Change* 2013: *The Physical Science Basis*. *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (T. F. Stocker, D. Qin, G. K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex et al., eds.) 867–952 10. Cambridge Univ. Press, Cambridge, United Kingdom and New York, NY, USA. https://doi.org/10.1017/CBO9781107415324.022

DELSOLE, T., TRENARY, L., YAN, X. and TIPPETT, M. K. (2019). Confidence intervals in optimal fingerprinting. *Clim. Dyn.* **52** 4111–4126.

ENGLE, R. F., LEDOIT, O. and WOLF, M. (2019). Large dynamic covariance matrices. *J. Bus. Econom. Statist.* **37** 363–375. MR3948411 https://doi.org/10.1080/07350015.2017.1345683

Key words and phrases. Measurement error, nonlinear shrinkage estimator, total least squares.

FULLER, W. A. (1980). Properties of some estimators for the errors-in-variables model. *Ann. Statist.* **8** 407–422. MR0560737

GLESER, L. J. (1981). Estimation in a multivariate "errors in variables" regression model: Large sample results. *Ann. Statist.* **9** 24–44. MR0600530

HANNART, A. (2016). Integrated optimal fingerprinting: Method description and illustration. *J. Climate* **29** 1977–1998.

HANNART, A., RIBES, A. and NAVEAU, P. (2014). Optimal fingerprinting under multiple sources of uncertainty. *Geophys. Res. Lett.* **41** 1261–1268.

HASSELMANN, K. (1997). Multi-pattern fingerprint method for detection and attribution of climate change. *Clim. Dyn.* **13** 601–611.

HEGERL, G. and ZWIERS, F. (2011). Use of models in detection and attribution of climate change. *Wiley Interdiscip. Rev.: Clim. Change* **2** 570–591.

HEGERL, G. C., VON STORCH, H., HASSELMANN, K., SANTER, B. D., CUBASCH, U. and JONES, P. D. (1996). Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. *J. Climate* **9** 2281–2306.

HEGERL, G. C., HOEGH-GULDBERG, O., CASASSA, G., HOERLING, M. P., KOVATS, R. S., PARMESAN, C., PIERCE, D. W. and STOTT, P. A. (2010). Good practice guidance paper on detection and attribution related to anthropogenic climate change. In *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Detection and Attribution of Anthropogenic Climate Change* (T. F. Stocker, C. B. Field, D. Qin, V. Barros, G. K. Plattner, M. Tignor, P. M. Midgley and K. L. Ebi, eds.) IPCC Working Group I Technical Support Unit, Univ. Bern, Bern, Switzerland.

JONES, G. S., STOTT, P. A. and CHRISTIDIS, N. (2013). Attribution of observed historical near–surface temperature variations to anthropogenic and natural causes using CMIP5 simulations. *J. Geophys. Res., Atmos.* **118** 4001–4024.

JONES, G. S., STOTT, P. A. and MITCHELL, J. F. B. (2016). Uncertainties in the attribution of greenhouse gas warming and implications for climate prediction. *J. Geophys. Res., Atmos.* **121** 6969–6992.

KATZFUSS, M., HAMMERLING, D. and SMITH, R. L. (2017). A Bayesian hierarchical model for climate change detection and attribution. *Geophys. Res. Lett.* **44** 5720–5728.

LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339 https://doi.org/10.1016/S0047-259X(03)00096-4

LEDOIT, O. and WOLF, M. (2017a). Direct nonlinear shrinkage estimation of large-dimensional covariance matrices Working Paper No. 264 Univ. Zurich, Department of Economics.

LEDOIT, O. and WOLF, M. (2017b). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *Rev. Financ. Stud.* **30** 4349–4388. https://doi.org/10.1093/rfs/hhx052

LEDOIT, O. and WOLF, M. (2018). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. *Bernoulli* **24** 3791–3832. MR3788189 https://doi.org/10.3150/17-BEJ979

LI, Y., CHEN, K. and YAN, J. (2020). dacc: Detection and Attribution of Climate Change. R package version 0.1-13.

LI, Y., CHEN, K., YAN, J. and ZHANG, X. (2021). Uncertainty in optimal fingerprinting is underestimated. *Environ. Res. Lett.* **16** 084043.

LI, Y., CHEN, K., YAN, J. and ZHANG, X. (2023). Supplement to "Regularized fingerprinting in detection and attribution of climate change with weight matrix optimizing the efficiency in scaling factor estimation." https://doi.org/10.1214/22-AOAS1624SUPPA, https://doi.org/10.1214/22-AOAS1624SUPPB

MORICE, C. P., KENNEDY, J. J., RAYNER, N. A. and JONES, P. D. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res., Atmos.* **117** D08101.

PEŠTA, M. (2013). Total least squares and bootstrapping with applications in calibration. *Statistics* **47** 966–991. MR3175728 https://doi.org/10.1080/02331888.2012.658806

RIBES, A., AZAÏS, J.-M. and PLANTON, S. (2009). Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate. *Clim. Dyn.* **33** 707–722.

RIBES, A., PLANTON, S. and TERRAY, L. (2013). Application of regularised optimal fingerprinting to attribution. Part I: Method, properties and idealised analysis. *Clim. Dyn.* **41** 2817–2836.

SCHURER, A., HEGERL, G., RIBES, A., POLSON, D., MORICE, C. and TETT, S. (2018). Estimating the transient climate response from observed warming. *J. Climate* **31** 8645–8663.

SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55** 331–339. MR1370408 https://doi.org/10.1006/jmva.1995.1083

ZHANG, X., ZWIERS, F. and STOTT, P. A. (2006). Multimodel multisignal climate change detection at regional scale. *J. Climate* **19** 4294–4307.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 240–268
https://doi.org/10.1214/22-AOAS1625

A MULTIVARIATE FREQUENCY-SEVERITY FRAMEWORK FOR HEALTHCARE DATA BREACHES

BY HONG SUN[1,a], MAOCHAO XU[2,b] AND PENG ZHAO[3,c]

[1]*School of Mathematics and Statistics, Lanzhou University,* [a]*sunh19@lzu.edu.cn*

[2]*Department of Mathematics, Illinois State University,* [b]*mxu2@ilstu.edu*

[3]*School of Mathematics and Statistics and RIMS, Jiangsu Provincial Key Laboratory of Educational Big Data Science and Engineering, Jiangsu Normal University,* [c]*zhaop@jsnu.edu.cn*

Data breaches in healthcare have become a substantial concern in recent years and cause millions of dollars in financial losses each year. It is fundamental for government regulators, insurance companies, and stakeholders to understand the breach frequency and the number of affected individuals in each state, as these are directly related to the federal Health Insurance Portability and Accountability Act (HIPAA) and state data breach laws. However, an obstacle to studying data breaches in healthcare is the lack of suitable statistical approaches. We develop a novel multivariate frequency-severity framework to analyze breach frequency and the number of affected individuals at the state level. A mixed effects model is developed to model the square root transformed frequency, and the log-gamma distribution is proposed to capture the skewness and heavy tail exhibited by the distribution of numbers of affected individuals. We further discover a positive nonlinear dependence between the transformed frequency and the log-transformed numbers of affected individuals (i.e., severity). In particular, we propose to use a D-vine copula to capture the multivariate dependence among conditional severities, given frequencies due to its inherent temporal structure and rich bivariate copula families. The rejection sampling technique is developed to simulate the predictive distributions. Both the in-sample and out-of-sample studies show that the proposed multivariate frequency-severity model that accommodates nonlinear dependence has satisfactory fitting and prediction performances.

REFERENCES

BALTAGI, B. (2008). *Econometric Analysis of Panel Data*. Wiley, New York.

BISOGNI, F. (2016). Proving limits of state data breach notification laws: Is a federal law the most adequate solution? *J. Inf. Policy* **6** 154–205.

BRECHMANN, E. and SCHEPSMEIER, U. (2013). Cdvine: Modeling dependence with c-and d-vine copulas in R. *J. Stat. Softw.* **52** 1–27.

BROOKS, M. E., KRISTENSEN, K., VAN BENTHEM, K. J., MAGNUSSON, A., BERG, C. W., NIELSEN, A., SKAUG, H. J., MACHLER, M. and BOLKER, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R J.* **9** 378–400.

BUCKMAN, J., HASHIM, M., WOUTERSEN, T. and BOCKSTEDT, J. (2018). Fool me twice: An analysis of repeat data breaches within firms. *SSRN Electron. J.* https://doi.org/10.2139/ssrn.3258599.

CUMMINS, J. D. and WILTBANK, L. J. (1983). Estimating the total claims distribution using multivariate frequency and severity distributions. *J. Risk Insur.* 377–403.

CZADO, C. (2019). *Analyzing Dependent Data with Vine Copulas*: *A Practical Guide with R*. *Lecture Notes in Statistics* **222**. Springer, Cham. MR3931334 https://doi.org/10.1007/978-3-030-13785-4

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

DOBSON, A. J. and BARNETT, A. G. (2018). *An Introduction to Generalized Linear Models*, 4th ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3890007

EDWARDS, B., HOFMEYR, S. and FORREST, S. (2016). Hype and heavy tails: A closer look at data breaches. *J. CyberSecurity* **2** 3–14.

Key words and phrases. Copula, data breach, heavy tail, multivariate dependence, score.

ELING, M. (2020). Cyber risk research in business and actuarial science. *Eur. Actuar. J.* **10** 303–333. MR4168360 https://doi.org/10.1007/s13385-020-00250-1

ELING, M. and JUNG, K. (2018). Copula approaches for modeling cross-sectional dependence of data breach losses. *Insurance Math. Econom.* **82** 167–180. MR3850616 https://doi.org/10.1016/j.insmatheco.2018.07.003

ELING, M. and WIRFS, J. (2019). What are the actual costs of cyber risk events? *European J. Oper. Res.* **272** 1109–1119.

FANG, Z., XU, M., XU, S. and HU, T. (2021). A framework for predicting data breach risk: Leveraging dependence to cope with sparsity. *IEEE Trans. Inform. Forensics Secur.* **1**.

FREES, E. W., DERRIG, R. A. and MEYERS, G. (2014). *Predictive Modeling Applications in Actuarial Science* **1**. Cambridge Univ. Press, Cambridge.

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

GONZÁLEZ-RIVERA, G., LEE, T.-H. and MISHRA, S. (2004). Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. *Int. J. Forecast.* **20** 629–645.

GRIMIT, E. P., GNEITING, T., BERROCAL, V. J. and JOHNSON, N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Q. J. R. Meteorol. Soc.* **132** 2925–2942.

HARTIG, F. (2020). DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models. Available at https://CRAN.R-project.org/package=DHARMa. R package version 0.3.3.0.

HHS, O. (2020). Breach portal: Notice to the secretary of HHS breach of unsecured protected health information. https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf.

HINKLEY, D. (1979). Predictive likelihood. *Ann. Statist.* **7** 718–728. MR0532237

HJ (2020a). Hipaa journal: Healthcare data breach statistics. https://www.hipaajournal.com/healthcare-data-breach-statistics/.

HJ (2020b). Hipaa journal: December 2019 healthcare data breach report. https://www.hipaajournal.com/december-2019-healthcare-data-breach-report/.

IBM (2020). Cost of a data breach report 2020. https://netdiligence.com/portfolio/cyber-claims-study/.

ITG (2020). Data breach notification laws by state. https://www.itgovernanceusa.com/data-breach-notification-laws.

JIANG, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. *Springer Series in Statistics*. Springer, New York. MR2308058

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

JOE, H. (2014). *Dependence Modeling with Copulas*. CRC Press/CRC, Boca Raton.

KAAS, R., GOOVAERTS, M., DHAENE, J. and DENUIT, M. (2008). *Modern Actuarial Risk Theory: Using R* **128**. Springer, Berlin.

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

LAMBERT, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** 1–14.

LEE, G. Y. and SHI, P. (2019). A dependent frequency-severity approach to modeling longitudinal insurance claims. *Insurance Math. Econom.* **87** 115–129. MR3949632 https://doi.org/10.1016/j.insmatheco.2019.04.004

MAILLART, T. and SORNETTE, D. (2010). Heavy-tailed distribution of cyber-risks. *Eur. Phys. J. B* **75** 357–364.

MAINDONALD, J. and BRAUN, J. (2006). *Data Analysis and Graphics Using R: An Example-Based Approach* **10**. Cambridge Univ. Press, Cambridge.

MATHESON, J. E. and WINKLER, R. L. (1976). Scoring rules for continuous probability distributions. *Manage. Sci.* **22** 1087–1096.

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

ROBERT, C. and CASELLA, G. (2013). *Monte Carlo Statistical Methods*. Springer, Berlin.

ROMANOSKY, S. (2016). Examining the costs and causes of cyber incidents. *J. CyberSecurity* **2** 121–135.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SHI, P., FENG, X. and IVANTSOVA, A. (2015). Dependent frequency-severity modeling of insurance claims. *Insurance Math. Econom.* **64** 417–428. MR3394665 https://doi.org/10.1016/j.insmatheco.2015.07.006

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

SHI, P. and ZHAO, Z. (2020). Regression for copula-linked compound distributions with applications in modeling aggregate insurance claims. *Ann. Appl. Stat.* **14** 357–380. MR4085097 https://doi.org/10.1214/19-AOAS1299

SKLAR, M. (1959). Fonctions de répartition à *n* dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **8** 229–231. MR0125600

SMITH, M. S. (2015). Copula modelling of dependence in multivariate time series. *Int. J. Forecast.* **31** 815–833.
SUN, H., XU, M. and ZHAO, P. (2020). Modeling malicious hacking data breach risks. *N. Am. Actuar. J.* 1–19.
SUN, H., XU, M. and ZHAO, P. (2023). Supplement to "A multivariate frequency-severity framework for healthcare data breaches." https://doi.org/10.1214/22-AOAS1625SUPPA, https://doi.org/10.1214/22-AOAS1625SUPPB
TAY, A. S. and WALLIS, K. F. (2000). Density forecasting: A survey. *J. Forecast.* **19** 235–254.
VERBEKE, G. (1997). Linear mixed models for longitudinal data. In *Linear Mixed Models in Practice* 63–153. Springer, Berlin.
WEST, B. T., WELCH, K. B. and GALECKI, A. T. (2014). *Linear Mixed Models: A Practical Guide Using Statistical Software*. CRC Press, Boca Raton.
WHEATLEY, S., MAILLART, T. and SORNETTE, D. (2016). The extreme risk of personal data breaches and the erosion of privacy. *Eur. Phys. J. B* **89** 7.
XU, M., SCHWEITZER, K. M., BATEMAN, R. M. and XU, S. (2018). Modeling and predicting cyber hacking breaches. *IEEE Trans. Inform. Forensics Secur.* **13** 2856–2871.
YARAGHI, N. and GOPAL, R. D. (2018). The role of HIPAA omnibus rules in reducing the frequency of medical data breaches: Insights from an empirical study. *Milbank Q.* **96** 144–166.
ZEILEIS, A., KLEIBER, C. and JACKMAN, S. (2008). Regression models for count data in R. *J. Stat. Softw.* **27** 1–25.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 269–293
https://doi.org/10.1214/22-AOAS1626

GENERALIZED THEME DICTIONARY MODELS FOR ASSOCIATION PATTERN DISCOVERY

BY YANG YANG[1,2,a] AND KE DENG[2,3,b]

[1]*School of Statistics and Data Science, LPMC and KLMDASR, Nankai University,* [a]*yangyang_nk@nankai.edu.cn*
[2]*Center for Statistical Science, Tsinghua University*
[3]*Department of Industrial Engineering, Tsinghua University,* [b]*kdeng@tsinghua.edu.cn*

Discovering association patterns of items from a collection of baskets composed of different items is an important problem in various fields. Assuming that each basket is composed of themes of items randomly sampled from a theme dictionary, the theme dictionary model provides a general framework to achieve efficient association pattern discovery with statistical inference. This paper extends the original theme dictionary model by allowing more than one category of items in a basket and only presence/absence of items is observed for each basket with all quantitative information missing. The extended models can solve a larger range of practical problems that cannot be handled by the original theme dictionary model. Both simulation studies and real data applications confirm the superiority of the proposed methods over the existing ones.

REFERENCES

AGRAWAL, R., IMIELINSKI, T. and SWAMI, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data* 207–216.

AGRAWAL, R. and SRIKANT, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the* 20*th International Conference on Very Large Data Bases* 487–499.

BLEI, D., ANDREW, Y. and JORDAN, M. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

BLEI, D. and LAFFERTY, J. (2006). Dynamic topic models. In *Proceedings of the* 23*rd International Conference on Machine Learning* 113–120.

BLEI, D. M. and LAFFERTY, J. D. (2007). A correlated topic model of Science. *Ann. Appl. Stat.* **1** 17–35. MR2393839 https://doi.org/10.1214/07-AOAS114

BOOTH, J. G. and HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 265–285.

BORGELT, C. (2003). Efficient implementations of apriori and eclat. In *Proceedings of the IEEE ICDM Workshop on Frequent Item Set Mining Implementations*.

BORGELT, C. (2004). Recursion pruning for the apriori algorithm. In *Proceedings of the IEEE ICDM Workshop on Frequent Item Set Mining Implementations*.

BORGELT, C. (2005). An implementation of the FP-growth algorithm. In *Proceedings of the* 1*st International Workshop on Open Source Data Mining* 1–5.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39** 1–38. MR0501537

DENG, K., GENG, Z. and LIU, J. S. (2014). Association pattern discovery via theme dictionary models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 319–347. MR3164869 https://doi.org/10.1111/rssb.12032

DENG, K., LIU, D., GAO, S. and GENG, Z. (2005). Structural learning of graphical models and its applications to traditional Chinese medicine. *Lecture Notes in Comput. Sci.* **3614** 362–367.

DOUCET, A., DE FREITAS, J. F. G. and GORDON, N. J. (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York.

FEARNHEAD, P. and CLIFFORD, P. (2003). On-line inference for hidden Markov models via particle filters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 887–899. MR2017876 https://doi.org/10.1111/1467-9868.00421

FENG, Y., WU, Z., ZHOU, X., ZHOU, Z. and FAN, W. (2006). Methodological review: Knowledge discovery in traditional Chinese medicine: State of the art and perspectives. *Artif. Intell. Med.* **38** 219–236.

Key words and phrases. Association rule mining, theme dictionary model, cross-category association patterns, missing data problem, Monte Carlo expectation-maximization algorithm.

GUPTA, N., MANGAL, N., TIWARI, K. and MITRA, P. (2006). Mining quantitative association rules in protein sequences. In *Data Mining* **3755**. Springer, Berlin.

HAN, J., PEI, J. and YIN, Y. (2000). Mining frequent patterns without candidate generation. *SIGMOD Rec*. **29** 1–12.

HE, P., DENG, K., LIU, Z., LIU, D., LIU, J. S. and GENG, Z. (2012). Discovering herbal functional groups of traditional Chinese medicine. *Stat*. *Med*. **31** 636–642. MR2900865 https://doi.org/10.1002/sim.4146

HUANG, Z., DONG, W., BATH, P., JI, L. and DUAN, H. (2015). On mining latent treatment patterns from electronic medical records. *Data Min*. *Knowl*. *Discov*. **29** 914–949. MR3353915 https://doi.org/10.1007/s10618-014-0381-y

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

LIU, J. S. and CHEN, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J*. *Amer*. *Statist*. *Assoc*. **93** 1032–1044. MR1649198 https://doi.org/10.2307/2669847

LIU, J. S., CHEN, R. and WONG, W. H. (1998). Rejection control and sequential importance sampling. *J*. *Amer*. *Statist*. *Assoc*. **93** 1022–1031. MR1649197 https://doi.org/10.2307/2669846

LU, H. (2020). Drug treatment options for the 2019-new coronavirus (2019-nCoV). *Biosci*. *Trends* **14** 69–71. https://doi.org/10.5582/bst.2020.01020

LUO, H., TANG, Q. L., SHANG, Y. X., LIANG, S. B., YANG, M., ROBINSON, N. and LIU, J. P. (2020). Can Chinese medicine be used for prevention of corona virus disease 2019 (COVID-19)? A review of historical classics, research evidence and current prevention programs. *Chin*. *J*. *Integr*. *Med*. **26** 243–250.

MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J*. *Amer*. *Statist*. *Assoc*. **92** 162–170. MR1436105 https://doi.org/10.2307/2291460

PIATETSKY-SHAPIRO, G. (1991). Discovery, analysis, and presentation of strong rules. *Knowl*. *Discov*. *Databases* 229–248.

RAJAK, A. and GUPTA, M. K. (2008). Association rule mining: Applications in various areas. In *Proceedings of International Conference on Data Management* 3–7.

ROSEN-ZVI, M., GRIFFITHS, T., STEYVERS, M. and SMYTH, P. (2004). The author-topic model for authors and documents. In *Proceedings of the* 20*th Conference on Uncertainty in Artificial Intelligence* 487–494.

RUIZ, F. J. R., ATHEY, S. and BLEI, D. M. (2020). SHOPPER: A probabilistic model of consumer choice with substitutes and complements. *Ann*. *Appl*. *Stat*. **14** 1–27. MR4085081 https://doi.org/10.1214/19-AOAS1265

WEBB, G. (2007). Discovering significant patterns. *Mach*. *Learn*. **68** 1–33.

YANG, Y. and DENG, K. (2023). Supplement to "Generalized Theme Dictionary Models for Association Pattern Discovery." https://doi.org/10.1214/22-AOAS1626SUPP

YANG, Y., LI, Q., LIU, Z., YE, F. and DENG, K. (2019). Understanding traditional Chinese medicine via statistical learning of expert-specific electronic medical records. *Quant*. *Biol*. **7** 201–232.

ZAKI, M. (2000). Scalable algorithms for association mining. *IEEE Trans*. *Knowl*. *Data Eng*. **12** 372–390.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 294–312
https://doi.org/10.1214/22-AOAS1627

ESTIMATING THE AVERAGE TREATMENT EFFECT IN RANDOMIZED CLINICAL TRIALS WITH ALL-OR-NONE COMPLIANCE

BY ZHIWEI ZHANG[1,a], ZONGHUI HU[2,b], DEAN FOLLMANN[2,c] AND LEI NIE[3,d]

[1]*Division of Cancer Treatment and Diagnosis, National Cancer Institute,* [a]*zhiwei.zhang@nih.gov*
[2]*Division of Clinical Research, National Institute of Allergy and Infectious Diseases,* [b]*huzo@niaid.nih.gov,* [c]*dfollmann@niaid.nih.gov*
[3]*Center for Drug Evaluation and Research, Food and Drug Administration,* [d]*lei.nie@fda.hhs.gov*

Noncompliance is a common intercurrent event in randomized clinical trials that raises important questions about analytical objectives and approaches. Motivated by the Multiple Risk Factor Intervention Trial (MRFIT), we consider how to estimate the average treatment effect (ATE) in randomized trials with all-or-none compliance. Confounding is a major challenge in estimating the ATE, and conventional methods for confounding adjustment typically require the assumption of no unmeasured confounders which may be difficult to justify. Using randomized treatment assignment as an instrumental variable, the ATE can be identified in the presence of unmeasured confounders under suitable assumptions, including an assumption that limits the effect-modifying activities of unmeasured confounders. We describe and compare several estimation methods based on different modeling assumptions. Some of these methods are able to incorporate information from auxiliary covariates for improved efficiency without introducing bias. The different methods are compared in a simulation study and applied to the MRFIT.

REFERENCES

ANGRIST, J. D. and FERNANDEZ-VAL, I. (2013). Extrapolate-ing: External validity and overidentification in the LATE framework. In *Advances in Economics and Econometrics*: *Theory and Applications*, *Vol. III*, *Econometrics* 401–434. Cambridge Univ. Press, Cambridge.

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

ARONOW, P. M. and CARNEGIE, A. (2013). Beyond LATE: Estimation of the average treatment effect with an instrumental variable. *Polit. Anal.* **21** 492–506.

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

D'AGOSTINO, R. B., VASAN, R. S., PENCINA, M. J., WOLF, P. A., COBAIN, M., MASSARO, J. M. and KANNEL, W. B. (2008). General cardiovascular risk profile for use in primary care: The framingham heart study. *Circulation* **117** 743–753.

FOLLMANN, D. A. (2000). On the effect of treatment among would-be treatment compliers: An analysis of the multiple risk factor intervention trial. *J. Amer. Statist. Assoc.* **95** 1101–1109. MR1821718 https://doi.org/10.2307/2669746

FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039 https://doi.org/10.1111/j.0006-341X.2002.00021.x

HARTWIG, E. P., WANG, L., SMITH, G. D. and DAVIES, N. M. (2020). Average causal effect estimation via instrumental variables: The no simultaneous heterogeneity assumption. https://arxiv.org/ftp/arxiv/papers/2010/2010.10017.pdf.

HE, J., STEPHENS-SHIELDS, A. and JOFFE, M. (2015). Structural nested mean models to estimate the effects of time-varying treatments on clustered outcomes. *Int. J. Biostat.* **11** 203–222. MR3427613 https://doi.org/10.1515/ijb-2014-0055

INTERNATIONAL CONFERENCE ON HARMONIZATION (2019). *Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials E9(R1)*. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf.

Key words and phrases. Complier-average treatment effect, effect modification, instrumental variable, noncompliance, principal stratification, unmeasured confounding.

MARK, S. D. and ROBINS, J. M. (2000a). A method for the analysis of randomized trials with compliance information: An application to the multiple risk factor intervention trial. *Control. Clin. Trials* **14** 79–97.

MARK, S. D. and ROBINS, J. M. (2000b). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Stat. Med.* **12** 1605–1628.

MULTIPLE RISK FACTOR INTERVENTION TRIAL (1977). Statistical design considerations in the NHLI Multiple Risk Factor Intervention Trial (MRFIT). *J. Chronic Dis.* **30** 261–275.

MULTIPLE RISK FACTOR INTERVENTION TRIAL (1982). Multiple risk factor intervention trial risk factor changes and mortality results. *J. Am. Med. Assoc.* **248** 1465–1477.

MULTIPLE RISK FACTOR INTERVENTION TRIAL (1996). Mortality after 16 years for participants randomized to the multiple risk factor intervention trial. *Circulation* **94** 946–951.

OGBURN, E. L., ROTNITZKY, A. and ROBINS, J. M. (2015). Doubly robust estimation of the local average treatment effect curve. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 373–396. MR3310531 https://doi.org/10.1111/rssb.12078

PEARL, J. (2011). Principal stratification—a goal or a tool? *Int. J. Biostat.* **7** Art. 20, 15. MR2787410 https://doi.org/10.2202/1557-4679.1322

ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect *Math. Model.* **7** 1393–1512. MR0877758 https://doi.org/10.1016/0270-0255(86)90088-6

ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412. MR1293185 https://doi.org/10.1080/03610929408831393

ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

WANG, L. and TCHETGEN TCHETGEN, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 531–550. MR3798877 https://doi.org/10.1111/rssb.12262

WANG, L., ZHANG, Y., RICHARDSON, T. S. and ROBINS, J. M. (2021). Estimation of local treatment effects under the binary instrumental variable model. *Biometrika* **108** 881–894. MR4341357 https://doi.org/10.1093/biomet/asab003

ZHANG, Z., HU, Z., FOLLMANN, D. and NIE, L. (2023). Supplement to "Estimating the average treatment effect in randomized clinical trials with all-or-none compliance." https://doi.org/10.1214/22-AOAS1627SUPPA, https://doi.org/10.1214/22-AOAS1627SUPPB

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 313–332
https://doi.org/10.1214/22-AOAS1628

MODEL SELECTION FOR MATERNAL HYPERTENSIVE DISORDERS WITH SYMMETRIC HIERARCHICAL DIRICHLET PROCESSES

BY BEATRICE FRANZOLINI[1,a], ANTONIO LIJOI[2,b] AND IGOR PRÜNSTER[2,c]

[1]*Singapore Institute for Clinical Sciences (SICS), Agency for Science, Technology and Research (A*STAR),* [a]*Beatrice_Franzolini@sics.a-star.edu.sg*

[2]*Department of Decision Sciences, Bocconi University,* [b]*antonio.lijoi@unibocconi.it,* [c]*igor.pruenster@unibocconi.it*

Hypertensive disorders of pregnancy occur in about 10% of pregnant women around the world. Though there is evidence that hypertension impacts maternal cardiac functions, the relation between hypertension and cardiac dysfunctions is only partially understood. The study of this relationship can be framed as a joint inferential problem on multiple populations, each corresponding to a different hypertensive disorder diagnosis, that combines multivariate information provided by a collection of cardiac function indexes. A Bayesian nonparametric approach seems particularly suited for this setup, and we demonstrate it on a dataset consisting of transthoracic echocardiography results of a cohort of Indian pregnant women. We are able to perform model selection, provide density estimates of cardiac function indexes and a latent clustering of patients: these readily interpretable inferential outputs allow to single out modified cardiac functions in hypertensive patients, compared to healthy subjects, and progressively increased alterations with the severity of the disorder. The analysis is based on a Bayesian nonparametric model that relies on a novel hierarchical structure, called symmetric hierarchical Dirichlet process. This is suitably designed so that the mean parameters are identified and used for model selection across populations, a penalization for multiplicity is enforced, and the presence of unobserved relevant factors is investigated through a latent clustering of subjects. Posterior inference relies on a suitable Markov chain Monte Carlo algorithm, and the model behaviour is also showcased on simulated data.

REFERENCES

AKSU, E., CUGLAN, B., TOK, A., CELIK, E., DOGANER, A., SOKMEN, A. and SOKMEN, G. (2021). Cardiac electrical and structural alterations in preeclampsia. *J. Matern.-Fetal Neonatal Med.* 1–10.

AMBROŽIC, J., LUCOVNIK, M., PROKŠELJ, K., TOPLIŠEK, J. and CVIJIC, M. (2020). Dynamic changes in cardiac function before and early postdelivery in women with severe preeclampsia. *J. Hypertens.* **38** 1367–1374.

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969

BELLAMY, L., CASAS, J.-P., HINGORANI, A. D. and WILLIAMS, D. J. (2007). Pre-eclampsia and risk of cardiovascular disease and cancer in later life: Systematic review and meta-analysis. *BMJ* **335** 974.

BERAHA, M., GUGLIELMI, A. and QUINTANA, F. A. (2021). The semi-hierarchical Dirichlet process and its application to clustering homogeneous distributions. *Bayesian Anal.* **16** 1187–1219. MR4381132 https://doi.org/10.1214/21-BA1278

CAMERLENGHI, F., DUNSON, D. B., LIJOI, A., PRÜNSTER, I. and RODRÍGUEZ, A. (2019a). Latent nested nonparametric priors (with discussion). *Bayesian Anal.* **14** 1303–1356. MR4044854 https://doi.org/10.1214/19-BA1169

CAMERLENGHI, F., LIJOI, A., ORBANZ, P. and PRÜNSTER, I. (2019b). Distribution theory for hierarchical processes. *Ann. Statist.* **47** 67–92. MR3909927 https://doi.org/10.1214/17-AOS1678

CHRISTENSEN, J. and MA, L. (2020). A Bayesian hierarchical model for related densities by using Pólya trees. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 127–153. MR4060979

Key words and phrases. Bayesian nonparametrics, clustering populations, Dirichlet process, hierarchical partitions, hierarchical process, hypertensive disorders of pregnancy, model based clustering.

CIFARELLI, D. M. and REGAZZINI, E. (1978). *Problemi Statistici Non Parametrici in Condizioni di Scambialbilita Parziale e Impiego di Medie Associative*. Quaderni Istituto Matematica Finanziaria, Torino.

CIPOLLI, W. III, HANSON, T. and MCLAIN, A. C. (2016). Bayesian nonparametric multiple testing. *Comput. Statist. Data Anal.* **101** 64–79. MR3504836 https://doi.org/10.1016/j.csda.2016.02.016

DAHL, D. B. and NEWTON, M. A. (2007). Multiple hypothesis testing by clustering treatment effects. *J. Amer. Statist. Assoc.* **102** 517–526. MR2325114 https://doi.org/10.1198/016214507000000211

DALAL, S. R. (1979a). Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions. *Stochastic Process. Appl.* **9** 99–107. MR0544719 https://doi.org/10.1016/0304-4149(79)90043-7

DALAL, S. R. (1979b). Nonparametric and robust Bayes estimation of location. In *Optimizing Methods in Statistics* (*Proc. Internat. Conf., Indian Inst. Tech., Bombay*, 1977) 141–166. Academic Press, New York. MR0541556

DAVIS, E. F., LAZDAM, M., LEWANDOWSKI, A. J., WORTON, S. A., KELLY, B., KENWORTHY, Y. et al. (2012). Cardiovascular risk factors in children and young adults born to preeclamptic pregnancies: A systematic review. *Pediatrics* **129** 1552–1561.

DE IORIO, M., MÜLLER, P., ROSNER, G. L. and MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99** 205–215. MR2054299 https://doi.org/10.1198/016214504000000205

DEMARTELLY, V. A., DREIXLER, J., TUNG, A., MUELLER, A., HEIMBERGER, S., FAZAL, A. A., NASEEM, H., LANG, R., KRUSE, E. et al. (2021). Long-term postpartum cardiac function and its association with preeclampsia. *J. Am. Heart Assoc.* **10** e018526.

DENTI, F., GUINDANI, M., LEISEN, F., LIJOI, A., WADSWORTH, W. D. and VANNUCCI, M. (2021). Two-group Poisson–Dirichlet mixtures for multiple testing. *Biometrics* **77** 622–633. MR4307660 https://doi.org/10.1111/biom.13314

DIACONIS, P. and FREEDMAN, D. (1986). On inconsistent Bayes estimates of location. *Ann. Statist.* **14** 68–87. MR0829556 https://doi.org/10.1214/aos/1176349843

DO, K.-A., MÜLLER, P. and TANG, F. (2005). A Bayesian mixture model for differential gene expression. *J. Roy. Statist. Soc. Ser. C* **54** 627–644. MR2137258 https://doi.org/10.1111/j.1467-9876.2005.05593.x

DOLEA, C. and ABOUZAHR, C. (2003). Global burden of hypertensive disorders of pregnancy in the year 2000 Technical report, GBD 2000 Working Paper, World Health Organization, Geneva.

DOSS, H. (1984). Bayesian estimation in the symmetric location problem. *Z. Wahrsch. Verw. Gebiete* **68** 127–147. MR0767797 https://doi.org/10.1007/BF00531774

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

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

FRANZOLINI, B., LIJOI, A. and PRÜNSTER, I. (2023). Supplement to "Model Selection for Maternal Hypertensive Disorders with Symmetric Hierarchical Dirichlet Processes." https://doi.org/10.1214/22-AOAS1628SUPPA, https://doi.org/10.1214/22-AOAS1628SUPPB

GARCIA-GONZALEZ, C., GEORGIOPOULOS, G., AZIM, S. A., MACAYA, F., KAMETAS, N., NIHOYANNOPOULOS, P., NICOLAIDES, K. H. and CHARAKIDA, M. (2020). Maternal cardiac assessment at 35 to 37 weeks improves prediction of development of preeclampsia. *Hypertens.* **76** 514–522.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*. CRC Press, Boca Raton, FL.

GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999). Consistent semiparametric Bayesian inference about a location parameter. *J. Statist. Plann. Inference* **77** 181–193. MR1687955 https://doi.org/10.1016/S0378-3758(98)00192-X

GOPALAN, R. and BERRY, D. A. (1998). Bayesian multiple comparisons using Dirichlet process priors. *J. Amer. Statist. Assoc.* **93** 1130–1139. MR1649207 https://doi.org/10.2307/2669856

GUINDANI, M., MÜLLER, P. and ZHANG, S. (2009). A Bayesian discovery procedure. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 905–925. MR2750250 https://doi.org/10.1111/j.1467-9868.2009.00714.x

GUTIÉRREZ, L., BARRIENTOS, A. F., GONZÁLEZ, J. and TAYLOR-RODRÍGUEZ, D. (2019). A Bayesian nonparametric multiple testing procedure for comparing several treatments against a control. *Bayesian Anal.* **14** 649–675. MR3959876 https://doi.org/10.1214/18-BA1122

HALL, M. E., GEORGE, E. M. and GRANGER, J. P. (2018). The heart during pregnancy. *Rev. Esp. Orientac.* **64** 1045–1050.

IGBERASE, G. and EBEIGBE, P. (2006). Eclampsia: Ten-years of experience in a rural tertiary hospital in the Niger Delta, Nigeria. *J. Obstet. Gynaecol.* **26** 414–417.

IGLESIAS, P. L., ORELLANA, Y. and QUINTANA, F. A. (2009). Nonparametric Bayesian modelling using skewed Dirichlet processes. *J. Statist. Plann. Inference* **139** 1203–1214. MR2479861 https://doi.org/10.1016/j.jspi.2008.07.009

LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20** 1260–1291. MR3217444 https://doi.org/10.3150/13-BEJ521

LIJOI, A., PRÜNSTER, I. and REBAUDO, G. (2022). Flexible clustering via hidden hierarchical Dirichlet priors. *Scand. J. Stat.*. https://doi.org/10.1111/sjos.12578

MACEACHERN, S. N. (2000). Dependent Dirichlet processes Technical Report, Department of Statistics, The Ohio State Univ.

MALIK, A., JEE, B. and GUPTA, S. K. (2019). Preeclampsia: Disease biology and burden, its management strategies with reference to India. *Pregnancy Hypertens.* **15** 23–31.

MARTIN, R. and TOKDAR, S. T. (2012). A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics* **13** 427–439.

MCCLURE, E. M., SALEEM, S., PASHA, O. and GOLDENBERG, R. L. (2009). Stillbirth in developing countries: A review of causes, risk factors and prevention strategies. *J. Matern.-Fetal Neonatal Med.* **22** 183–190.

MOSER, S., RODRÍGUEZ, A. and LOFLAND, C. L. (2021). Multiple ideal points: Revealed preferences in different domains. *Polit. Anal.* **29** 139–166.

MULIERE, P. and PETRONE, S. (1993). A Bayesian predictive approach to sequential search for an optimal dose: Parametric and nonparametric models. *J. Italian Stat. Soc.* **2** 349–364.

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 https://doi.org/10.2307/1390653

PEDERSEN, S. S., VON KÄNEL, R., TULLY, P. J. and DENOLLET, J. (2017). Psychosocial perspectives in cardiovascular disease. *Eur. J. Prev. Cardiol.* **24** 108–115.

RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2008). The nested Dirichlet process. *J. Amer. Statist. Assoc.* **103** 1131–1144. MR2528831 https://doi.org/10.1198/016214508000000553

SCOTT, J. G. and BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference* **136** 2144–2162. MR2235051 https://doi.org/10.1016/j.jspi.2005.08.031

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450 https://doi.org/10.1214/10-AOS792

SHAH, A., FAWOLE, B., M'IMUNYA, J. M., AMOKRANE, F., NAFIOU, I., WOLOMBY, J.-J. et al. (2009). Cesarean delivery outcomes from the WHO global survey on maternal and perinatal health in Africa. *Int. J. Gynecol. Obstet.* **107** 191–197.

SORIANO, J. and MA, L. (2017). Probabilistic multi-resolution scanning for two-sample differences. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 547–572. MR3611759 https://doi.org/10.1111/rssb.12180

TATAPUDI, R. and PASUMARTHY, L. R. (2017a). Data for: Maternal cardiac function in gestational hypertension, mild and severe preeclampsia and normal pregnancy: A comparative study. Available at https://data.mendeley.com/datasets/d72zr4xggx/1. https://doi.org/10.17632/d72zr4xggx.1 Licensed under a Creative Commons Attribution 4.0 International licence.

TATAPUDI, R. and PASUMARTHY, L. R. (2017b). Maternal cardiac function in gestational hypertension, mild and severe preeclampsia and normal pregnancy: A comparative study. *Pregnancy Hypertens.* **10** 238–241.

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

TIMOKHINA, E., KUZMINA, T., STRIZHAKOV, A., PITSKHELAURI, E., IGNATKO, I. and BELOUSOVA, V. (2019). Maternal cardiac function after normal delivery, preeclampsia, and eclampsia: A prospective study. *J. Pregnancy* **2019** 2090–2727.

ZUANETTI, D. A., MÜLLER, P., ZHU, Y., YANG, S. and JI, Y. (2018). Clustering distributions with the marginalized nested Dirichlet process. *Biometrics* **74** 584–594. MR3825345 https://doi.org/10.1111/biom.12778

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 333–356
https://doi.org/10.1214/22-AOAS1630

BAYESIAN NON-HOMOGENEOUS HIDDEN MARKOV MODEL WITH VARIABLE SELECTION FOR INVESTIGATING DRIVERS OF SEIZURE RISK CYCLING

BY EMILY T. WANG[1,a], SHARON CHIANG[2,c], ZULFI HANEEF[3,e], VIKRAM R. RAO[2,d], ROBERT MOSS[4,f] AND MARINA VANNUCCI[1,b]

[1]*Department of Statistics, Rice University,* [a]*emilywang@rice.edu,* [b]*marina@rice.edu*
[2]*Department of Neurology, University of California, San Francisco,* [c]*sharon.chiang@ucsf.edu,* [d]*vikram.rao@ucsf.edu*
[3]*Department of Neurology, Baylor College of Medicine,* [e]*zulfi.haneef@bcm.edu*
[4]*Seizure Tracker, LLC,* [f]*rob@seizuretracker.com*

A major issue in the clinical management of epilepsy is the unpredictability of seizures. Yet, traditional approaches to seizure forecasting and risk assessment in epilepsy rely heavily on raw seizure frequencies which are a stochastic measurement of seizure risk. We consider a Bayesian nonhomogeneous hidden Markov model for unsupervised clustering of zero-inflated seizure count data. The proposed model allows for a probabilistic estimate of the sequence of seizure risk states at the individual level. It also offers significant improvement over prior approaches by incorporating a variable selection prior for the identification of clinical covariates that drive seizure risk changes and accommodating highly granular data. For inference, we implement an efficient sampler that employs stochastic search and data augmentation techniques. We evaluate model performance on simulated seizure count data. We then demonstrate the clinical utility of the proposed model by analyzing daily seizure count data from 133 patients with Dravet syndrome collected through the *Seizure Tracker*^{TM} system, a patient-reported electronic seizure diary. We report on the dynamics of seizure risk cycling, including validation of several known pharmacologic relationships. We also uncover novel findings characterizing the presence and volatility of risk states in Dravet syndrome which may directly inform counseling to reduce the unpredictability of seizures for patients with this devastating cause of epilepsy.

REFERENCES

ARTHURS, S., ZAVERI, H. P., FREI, M. G. and OSORIO, I. (2010). Patient and caregiver perspectives on seizure prediction. *Epilepsy Behav.* **19** 474–477.

AYUB, N., CHIANG, S., MOSS, R. and GOLDENHOLZ, D. (2020). Natural history of generalized motor seizures: A retrospective analysis. *Seizure* **80** 109–112. https://doi.org/10.1016/j.seizure.2020.05.019

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192 https://doi.org/10.1214/009053604000000238

BAUD, M. O., KLEEN, J. K., MIRRO, E. A., ANDRECHAK, J. C., KING-STEPHENS, D., CHANG, E. F. and RAO, V. R. (2018). Multi-day rhythms modulate seizure risk in epilepsy. *Nat. Commun.* **9** 88. https://doi.org/10.1038/s41467-017-02577-y

BAUD, M. O., PROIX, T., RAO, V. R. and SCHINDLER, K. (2020). Chance and risk in epilepsy. *Curr. Opin. Neurol.* **33** 163–172. https://doi.org/10.1097/WCO.0000000000000798

BROWN, P. J., VANNUCCI, M. and FEARN, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 627–641. MR1626005 https://doi.org/10.1111/1467-9868.00144

CANALE, A. and DUNSON, D. B. (2011). Bayesian kernel mixtures for counts. *J. Amer. Statist. Assoc.* **106** 1528–1539. MR2896854 https://doi.org/10.1198/jasa.2011.tm10552

CHIANG, S., VANNUCCI, M., GOLDENHOLZ, D. M., MOSS, R. and STERN, J. M. (2018). Epilepsy as a dynamic disease: A Bayesian model for differentiating seizure risk from natural variability. *Epilepsia Open* **3** 236–246. https://doi.org/10.1002/epi4.12112

Key words and phrases. Bayesian inference, count data, Dravet syndrome, epilepsy, hidden Markov models, Markov chain Monte Carlo, seizure risk, zero-inflation.

CHIANG, S., GOLDENHOLZ, D. M., MOSS, R., RAO, V. R., HANEEF, Z., THEODORE, W. H., KLEEN, J. K., GAVVALA, J., VANNUCCI, M. et al. (2020). Prospective validation study of an epilepsy seizure risk system for outpatient evaluation. *Epilepsia* **61** 29–38. https://doi.org/10.1111/epi.16397

CHIANG, S., KHAMBHATI, A. N., WANG, E. T., VANNUCCI, M., CHANG, E. F. and RAO, V. R. (2021). Evidence of state-dependence in the effectiveness of responsive neurostimulation for seizure modulation. *Brain Stimul*. **14** 366–375.

COOPER, M. S., MCINTOSH, A., CROMPTON, D. E., MCMAHON, J. M., SCHNEIDER, A., FARRELL, K., GANESAN, V., GILL, D., KIVITY, S. et al. (2016). Mortality in Dravet syndrome. *Epilepsy Res*. **128** 43–47. https://doi.org/10.1016/j.eplepsyres.2016.10.006

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

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist*. *Sinica* **7** 339–373.

GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, 4 (*Peñíscola*, 1991) 169–193. Oxford Univ. Press, New York. MR1380276

GOLDENHOLZ, D. M., GOLDENHOLZ, S. R., MOSS, R., FRENCH, J., LOWENSTEIN, D., KUZNIECKY, R., HAUT, S., CRISTOFARO, S., DETYNIECKI, K. et al. (2018). Is seizure frequency variance a predictable quantity? *Ann*. *Clin*. *Transl*. *Neurology* **5** 201–207. https://doi.org/10.1002/acn3.519

HARDEN, C., TOMSON, T., GLOSS, D., BUCHHALTER, J., CROSS, J. H., DONNER, E., FRENCH, J. A., GIL-NAGEL, A., HESDORFFER, D. C. et al. (2017). Practice guideline summary: Sudden unexpected death in epilepsy incidence rates and risk factors: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology and the American Epilepsy Society. *Epilepsy Curr.* **17** 180–187.

HAUT, S. R., HALL, C. B., MASUR, J. and LIPTON, R. B. (2007). Seizure occurrence: Precipitants and prediction. *Neurology* **69** 1905–1910.

HOLMES, C. C. and HELD, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal*. **1** 145–168. MR2227368 https://doi.org/10.1214/06-BA105

HOLSCLAW, T., GREENE, A. M., ROBERTSON, A. W. and SMYTH, P. (2017). Bayesian nonhomogeneous Markov models via Pólya-gamma data augmentation with applications to rainfall modeling. *Ann*. *Appl*. *Stat*. **11** 393–426. MR3634329 https://doi.org/10.1214/16-AOAS1009

HUBIN, A. (2019). An adaptive simulated annealing EM algorithm for inference on non-homogeneous hidden Markov models. In *Proceedings of the International Conference on Artificial Intelligence*, *Information Processing and Cloud Computing—AIIPCC'*19. https://doi.org/10.1145/3371425.3371641

KAROLY, P. J., GOLDENHOLZ, D. M., FREESTONE, D. R., MOSS, R. E., GRAYDEN, D. B., THEODORE, W. H. and COOK, M. J. (2018). Circadian and circaseptan rhythms in human epilepsy: A retrospective cohort study. *Lancet Neurol*. **17** 977–985. https://doi.org/10.1016/S1474-4422(18)30274-6

LEGUIA, M. G., ANDRZEJAK, R. G., RUMMEL, C., FAN, J. M., MIRRO, E. A., TCHENG, T. K., RAO, V. R. and BAUD, M. O. (2021). Seizure cycles in focal epilepsy. *JAMA Neurology*.

MUENZ, L. R. and RUBINSTEIN, L. V. (1985). Markov models for covariate dependence of binary sequences. *Biometrics* **41** 91–101. MR0793435 https://doi.org/10.2307/2530646

NAIMO, G. D., GUARNACCIA, M., SPROVIERI, T., UNGARO, C., CONFORTI, F. L., ANDÒ, S. and CAVALLARO, S. (2019). A systems biology approach for personalized medicine in refractory epilepsy. *Int*. *J*. *Mol*. *Sci*. **20** 3717. https://doi.org/10.3390/ijms20153717

PANAYIOTOPOULOS, C. P. (2005). Chapter 4, Principles of Therapy in Epilepsies. In *The Epilepsies*: *Seizures*, *Syndromes and Management* (C. P. Panayiotopoulos, ed.) Bladon Medical Publishing, Oxfordshire, UK.

PILLOW, J. W. and SCOTT, J. (2012). Fully Bayesian inference for neural models with negative-binomial spiking. In *Advances in Neural Information Processing Systems* 25 (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) 1898–1906. Curran Associates, Red Hook.

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

PRIVITERA, M. D., WELTY, T. E., GIDAL, B. E., DIAZ, F. J., KREBILL, R., SZAFLARSKI, J. P., DWORETZKY, B. A., POLLARD, J. R., ELDER JR., E. J. et al. (2016). Generic-to-generic lamotrigine switches in people with epilepsy: The randomised controlled EQUIGEN trial. *Lancet Neurol*. **15** 365–372.

PROIX, T., TRUCCOLO, W., LEGUIA, M. G., TCHENG, T. K., KING-STEPHENS, D., RAO, V. R. and BAUD, M. O. (2021). Forecasting seizure risk in adults with focal epilepsy: A development and validation study. *Lancet Neurol*. **20** 127–135. https://doi.org/10.1016/S1474-4422(20)30396-3

QUENOUILLE, M. H. (1949). A relation between the logarithmic, Poisson, and negative binomial series. *Biometrics* **5** 162–164. MR0030163 https://doi.org/10.2307/3001917

RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Readings in Speech Recognition* **77** 257–286. https://doi.org/10.1016/b978-0-08-051584-7.50027-9

RAO, V. R., LEGUIA, M. G., TCHENG, T. K. and BAUD, M. O. (2020). Cues for seizure timing. *Epilepsia*.

SAVITSKY, T., VANNUCCI, M. and SHA, N. (2011). Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statist. Sci.* **26** 130–149. MR2849913 https://doi.org/10.1214/11-STS354

SCOTT, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *J. Amer. Statist. Assoc.* **97** 337–351. MR1963393 https://doi.org/10.1198/016214502753479464

STINGO, F. C., CHEN, Y. A., VANNUCCI, M., BARRIER, M. and MIRKES, P. E. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. *Ann. Appl. Stat.* **4** 2024–2048. MR2829945 https://doi.org/10.1214/10-AOAS360

THARAYIL, J. J., CHIANG, S., MOSS, R., STERN, J. M., THEODORE, W. H. and GOLDENHOLZ, D. M. (2017). A big data approach to the development of mixed-effects models for seizure count data. *Epilepsia* **58** 835–844. https://doi.org/10.1111/epi.13727

VOSSLER, D. G., ANDERSON, G. D. and BAINBRIDGE, J. (2016). AES position statement on generic substitution of antiepileptic drugs. *Epilepsy Curr.* **16** 209–211.

WALCZAK, T. S., LEPPIK, I. E., D'AMELIO, M., RARICK, J., SO, E., AHMAN, P., RUGGLES, K., CASCINO, G. D., ANNEGERS, J. et al. (2001). Incidence and risk factors in sudden unexpected death in epilepsy: A prospective cohort study. *Neurology* **56** 519–525.

WANG, E. T., CHIANG, S., HANEEF, Z., RAO, V. R., MOSS, R. and VANNUCCI, M. (2023). Supplement to "Bayesian non-homogeneous hidden Markov model with variable selection for investigating drivers of seizure risk cycling." https://doi.org/10.1214/22-AOAS1630SUPPA, https://doi.org/10.1214/22-AOAS1630SUPPB

WHO, ILAE and IBE (2019). Epilepsy: A public health imperative.

WU, Y. W., SULLIVAN, J., MCDANIEL, S. S., MEISLER, M. H., WALSH, E. M., LI, S. X. and KUZNIEWICZ, M. W. (2015). Incidence of Dravet syndrome in a US population. *Pediatrics* **136** e1310–e1315.

YOSHITOMI, S., TAKAHASHI, Y., YAMAGUCHI, T., IMAI, K., ISHII, A., HIROSE, S. and INOUE, Y. (2019). Efficacy and tolerability of perampanel in pediatric patients with Dravet syndrome. *Epilepsy Res.* **154** 34–38.

ZHOU, M. and CARIN, L. (2015). Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 307–320.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 357–377
https://doi.org/10.1214/22-AOAS1631

MODELING CELL POPULATIONS MEASURED BY FLOW CYTOMETRY WITH COVARIATES USING SPARSE MIXTURE OF REGRESSIONS

BY SANGWON HYUN[1,a], MATTIAS ROLF CAPE[2,c], FRANCOIS RIBALET[2,d] AND JACOB BIEN[1,b]

[1]*Department of Data Sciences and Operations, University of Southern California,* [a]*sangwonh@ucsc.edu,* [b]*jbien@usc.edu*
[2]*School of Oceanography, University of Washington,* [c]*mcape@uw.edu,* [d]*ribalet@uw.edu*

The ocean is filled with microscopic microalgae, called phytoplankton, which together are responsible for as much photosynthesis as all plants on land combined. Our ability to predict their response to the warming ocean relies on understanding how the dynamics of phytoplankton populations is influenced by changes in environmental conditions. One powerful technique to study the dynamics of phytoplankton is flow cytometry which measures the optical properties of thousands of individual cells per second. Today, oceanographers are able to collect flow cytometry data in real time onboard a moving ship, providing them with fine-scale resolution of the distribution of phytoplankton across thousands of kilometers. One of the current challenges is to understand how these small- and large-scale variations relate to environmental conditions, such as nutrient availability, temperature, light and ocean currents. In this paper we propose a novel sparse mixture of multivariate regressions model to estimate the time-varying phytoplankton subpopulations while simultaneously identifying the specific environmental covariates that are predictive of the observed changes to these subpopulations. We demonstrate the usefulness and interpretability of the approach using both synthetic data and real observations collected on an oceanographic cruise conducted in the northeast Pacific in the spring of 2017.

REFERENCES

AGHAEEPOUR, N., FINAK, G., CONSORTIUM, F., CONSORTIUM, D. R. E. A. M., HOOS, H., MOSMANN, T. R., BRINKMAN, R., GOTTARDO, R. and SCHEUERMANN, R. H. (2013). Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **10** 228–238.

ASHKEZARI, M. D., HAGEN, N. R., DENHOLTZ, M., NEANG, A., BURNS, T. C., MORALES, R. L., LEE, C. P., HILL, C. N. and ARMBRUST, E. V. (2021). Simons collaborative marine atlas project (Simons CMAP): An open-source portal to share, visualize and analyze ocean data. *BioRxiv*.

BERUBE, P. M., BILLER, S. J., KENT, A. G., BERTA-THOMPSON, J. W., ROGGENSACK, S. E., ROACHE-JOHNSON, K. H., ACKERMAN, M., MOORE, L. R., MEISEL, J. D. et al. (2015). Physiology and evolution of nitrate acquisition in prochlorococcus. *ISME J.* **9** 1195–1207.

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.

BOYER, T. P., ANTONOV, J. I., BARANOVA, O. K., GARCIA, H. E., JOHNSON, D. R., MISHONOV, A. V., O'BRIEN, T. D., SEIDOV, D., SMOLYAR, I. I. et al. (2013). World ocean database 2013.

BROWN, M. and WITTWER, C. (2000). Flow cytometry: Principles and clinical applications in hematology. *Clin. Chem.* **46** 1221–1229.

DUBELAAR, G. B. J., GERRITZEN, P. L., BEEKER, A. E. R., JONKER, R. R. and TANGEN, K. (1999). Design and first results of CytoBuoy: A wireless flow cytometer for in situ analysis of marine and fresh waters. *Cytometry* **37** 247–254.

FIELD, C. B., BEHRENFELD, M. J., RANDERSON, J. T. and FALKOWSKI, P. (1998). Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281** 237–240.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

Key words and phrases. Mixture of regressions, expectation-maximization, flow cytometry, sparse regression, ocean, microbiome, phytoplankton, clustering, gating, alternating direction method of multipliers.

GEDYE, C. A., HUSSAIN, A., PATERSON, J., SMRKE, A., SAINI, H., SIRSKYJ, D., PEREIRA, K., LOBO, N., STEWART, J. et al. (2014). Cell surface profiling using high-throughput flow cytometry: A platform for biomarker discovery and analysis of cellular heterogeneity. *PLoS ONE* **9** e105602.

GRADOVILLE, M. R., FARNELID, H., WHITE, A. E., TURK-KUBO, K. A., STEWART, B., RIBALET, F., FERRÓN, S., PINEDO-GONZALEZ, P., ARMBRUST, E. V. et al. (2020). Latitudinal constraints on the abundance and activity of the cyanobacterium UCYN-A and other marine diazotrophs in the North Pacific. *Limnol. Oceanogr.* **65** 1858–1875.

GRANT, M. and BOYD, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx.

GRÜN, B. and LEISCH, F. (2008). FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *J. Stat. Softw.* **28** 1–35.

HAHNE, F., LEMEUR, N., BRINKMAN, R. R., ELLIS, B., HAALAND, P., SARKAR, D., SPIDLEN, J., STRAIN, E. and GENTLEMAN, R. (2009). FlowCore: A bioconductor package for high throughput flow cytometry. *BMC Bioinform.* **10** 106. https://doi.org/10.1186/1471-2105-10-106

HYRKAS, J., CLAYTON, S., RIBALET, F., HALPERIN, D., VIRGINIA ARMBRUST, E. and HOWE, B. (2015). Scalable clustering algorithms for continuous environmental flow cytometry. *Bioinformatics* **32** 417–423.

HYUN, S., MISHRA, A., MÜLLER, C. and BIEN, J. (2020). R package for CMAP Access. [Online; accessed 19-Dec-2019].

HYUN, S., ROLF CAPE, M., RIBALET, F. and BIEN, J. (2023). Supplement to "Modeling Cell Populations Measured By Flow Cytometry With Covariates Using Sparse Mixture of Regressions." https://doi.org/10.1214/22-AOAS1631SUPP

JORDAN, M. I. and JACOBS, R. A. (1993). Hierarchical mixtures of experts and the EM algorithm. In *Proceedings of* 1993 *International Conference on Neural Networks* (*IJCNN*-93-*Nagoya*, *Japan*).

KHALILI, A. and CHEN, J. (2007). Variable selection in finite mixture of regression models. *J. Amer. Statist. Assoc.* **102** 1025–1038. MR2411662 https://doi.org/10.1198/016214507000000590

MALEK, M., TAGHIYAR, M. J., CHONG, L., FINAK, G., GOTTARDO, R. and BRINKMAN, R. R. (2015). FlowDensity: Reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics* **31** 606–607.

MARAÑÓN, E. (2015). Cell size as a key determinant of phytoplankton metabolism and community structure. *Annu. Rev. Mar. Sci.* **7** 241–264.

MCLACHLAN, G. J. and PEEL, D. (2006). *Finite Mixture Models*. Wiley, New York.

MOORE, L. R., POST, A. F., ROCAP, G. and CHISHOLM, S. W. (2002). Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Limnol. Oceanogr.* **47** 989–996.

OLSON, R. J., SHALAPYONOK, A. and SOSIK, H. M. (2003). An automated submersible flow cytometer for analyzing pico- and nanophytoplankton: FlowCytobot. *Deep-Sea Res.*, *Part* 1, *Oceanogr. Res. Pap.* **50** 301–315.

PAN, W. and SHEN, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* **8** 1145–1164.

PROPS, R., MONSIEURS, P., MYSARA, M., CLEMENT, L. and BOON, N. (2016). Measuring the biodiversity of microbial communities by flow cytometry. *Methods Ecol. Evol.* **7** 1376–1385.

RIBALET, F., SWALWELL, J., CLAYTON, S., JIMÉNEZ, V., SUDEK, S., LIN, Y., JOHNSON, Z. I., WORDEN, A. Z. and ARMBRUST, E. V. (2015). Light-driven synchrony of Prochlorococcus growth and mortality in the subtropical Pacific gyre. *Proc. Natl. Acad. Sci. USA* **112** 8008–8012.

RIBALET, F., BERTHIAUME, C., HYNES, A., SWALWELL, J., CARLSON, M., CLAYTON, S., HENNON, G., POIRIER, C., SHIMABUKURO, E. et al. (2019). SeaFlow data v1, high-resolution abundance, size and biomass of small phytoplankton in the North Pacific. *Sci. Data* **6** 277. https://doi.org/10.1038/s41597-019-0292-2

SOSIK, H. M., OLSON, R. J. and ARMBRUST, E. V. (2010). Flow cytometry in phytoplankton research. In *Chlorophyll a Fluorescence in Aquatic Sciences*: *Methods and Applications* 171–185. Springer, Berlin.

SOSIK, H. M., OLSON, R. J., NEUBERT, M. G., SHALAPYONOK, A. and SOLOW, A. R. (2003). Growth rates of coastal phytoplankton from time-series measurements with a submersible flow cytometer. *Limnol. Oceanogr.* **48** 1756–1765.

STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ_1-penalization for mixture regression models. *TEST* **19** 209–256. MR2677722 https://doi.org/10.1007/s11749-010-0197-z

SWALWELL, J. E., RIBALET, F. and ARMBRUST, E. V. (2011). SeaFlow: A novel underway flow-cytometer for continuous observations of phytoplankton in the ocean. *Limnol. Oceanogr.*, *Methods* **9** 466–477.

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

VAULOT, D. and MARIE, D. (1999). Diel variability of photosynthetic picoplankton in the equatorial Pacific. *J. Geophys. Res.*, *Oceans* **104** 3297–3310.

VERSCHOOR, C. P., LELIC, A., BRAMSON, J. L. and BOWDISH, D. M. E. (2015). An introduction to automated flow cytometry gating tools and their implementation. *Front. Immunol.* **6** 380.

WANG, P., PUTERMAN, M. L., COCKBURN, I. and LE, N. (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics* **52** 381–400.

XU, L. and JORDAN, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Comput.* **8** 129–151.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 378–402
https://doi.org/10.1214/22-AOAS1632

GENERALIZED FIDUCIAL FACTOR: AN ALTERNATIVE TO THE BAYES FACTOR FOR FORENSIC IDENTIFICATION OF SOURCE PROBLEMS

BY JONATHAN P. WILLIAMS[1,a], DANICA M. OMMEN[2,b] AND JAN HANNIG[3,c]

[1]*Department of Statistics, North Carolina State University,* [a]*jwilli27@ncsu.edu*

[2]*Department of Statistics, Iowa State University,* [b]*dmommen@iastate.edu*

[3]*Department of Statistics and Operations Research, UNC at Chapel Hill and National Institute of Standards and Technology,* [c]*jan.hannig@unc.edu*

One formulation of forensic identification of source problems is to determine the source of trace evidence, for instance, glass fragments found on a suspect for a crime. The current state of the science is to compute a Bayes factor comparing the marginal distribution of measurements of trace evidence under two competing propositions for whether or not the unknown source evidence originated from a specific source. The obvious problem with such an approach is the ability to tailor the prior distributions (placed on the features/parameters of the statistical model for the measurements of trace evidence) in favor of the defense or prosecution which is further complicated by the fact that the typical number of measurements of trace evidence is typically sufficiently small that prior choice/specification has a strong influence on the value of the Bayes factor. To remedy this problem of prior specification and choice, we develop an alternative to the Bayes factor, within the framework of generalized fiducial inference, that we term a *generalized fiducial factor*. Furthermore, we demonstrate empirically, on synthetic and real Netherlands Forensic Institute casework data, deficiencies in Bayes factor and classical/frequentist likelihood ratio approaches.

REFERENCES

AITKEN, C. G. G. and LUCY, D. (2004). Evaluation of trace evidence in the form of multivariate data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **53** 109–122. MR2037883 https://doi.org/10.1046/j.0035-9254.2003.05271.x

BERGER, C. E. H. and SLOOTEN, K. (2016). The LR does not exist. *Sci. Justice* **56** 388–391. https://doi.org/10.1016/j.scijus.2016.06.005

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37** 905–938. MR2502655 https://doi.org/10.1214/07-AOS587

BERGER, J. O. and PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122. MR1394065 https://doi.org/10.2307/2291387

BERGER, J. O., PERICCHI, L. R., GHOSH, J., SAMANTA, T., DE SANTIS, F., BERGER, J. and PERICCHI, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In *Lecture Notes—Monograph Series* 135–207.

BERNARDO, J.-M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B* **41** 113–147. MR0547240

BIEDERMANN, A., BOZZA, S., TARONI, F. and AITKEN, C. G. G. (2016). Reframing the debate: A question of probability, not of likelihood ratio. *Sci. Justice* **56** 392–396.

BOLCK, A., WEYERMANN, C., DUJOURDY, L., ESSEIVA, P. and VAN DEN BERG, J. (2009). Different likelihood ratio approaches to evaluate the strength of MDMA tablet comparisons. *Forensic Sci. Int.* **191** 42–51.

CUI, Y. and HANNIG, J. (2019). Nonparametric generalized fiducial inference for survival functions under censoring. *Biometrika* **106** 501–518. MR3992384 https://doi.org/10.1093/biomet/asz016

DETTMAN, J. R., CASSABAUM, A. A., SAUNDERS, C. P., SNYDER, D. L. and BUSCAGLIA, J. (2014). Forensic discrimination of copper wire using trace element concentrations. *Anal. Chem.* **86** 8176–8182. https://doi.org/10.1021/ac5013514.

DICICCIO, T. J., KASS, R. E., RAFTERY, A. and WASSERMAN, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.* **92** 903–915. MR1482122 https://doi.org/10.2307/2965554

Key words and phrases. Bayes factor, generalized fiducial inference, likelihood ratio.

EGLI, N. M., CHAMPOD, C. and MARGOT, P. (2006). Evidence evaluation in fingerprint comparison and automated fingerprint identification systems—modeling between finger variability. *Forensic Sci. Int.* **176** 189–195.

ENFSI (2015). Enfsi guideline for evaluative reporting in forensic science. http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.

EVETT, I. W. (1977). The interpretation of refractive index measurements. *J. Forensic Sci.* **9** 209–217.

EVETT, I. W. (1986). A Bayesian approach to the problem of interpreting glass evidence in forensic science casework. *J. - Forensic Sci. Soc.* **26** 3–18.

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

GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185. MR1647507 https://doi.org/10.1214/ss/1028905934

GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2** 1360–1383. MR2655663 https://doi.org/10.1214/08-AOAS191

GONZALEZ-RODRIGUEZ, J., DRYGAJLO, A., RAMOS-CASTRO, D., GARCIA-GOMAR, M. and ORTEGA-GARCIA, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Comput. Speech Lang.* **20** 331–355.

GONZALEZ-RODRIGUEZ, J., FIERREZ-AGUILAR, J., RAMOS-CASTRO, D. and ORTEGA-GARCIA, J. (2005). Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems. *Forensic Sci. Int.* **155** 126–140.

GROVE, D. M. (1980). The interpretation of forensic evidence using a likelihood ratio. *Biometrika* **67** 243–246. MR0570530 https://doi.org/10.1093/biomet/67.1.243

HANNIG, J. and IYER, H. (2022). Testing for calibration discrepancy of reported likelihood ratios in forensic science. *J. Roy. Statist. Soc. Ser. A* **185** 267–301. MR4384306

HANNIG, J., IYER, H., LAI, R. C. S. and LEE, T. C. M. (2016). Generalized fiducial inference: A review and new results. *J. Amer. Statist. Assoc.* **111** 1346–1361. MR3561954 https://doi.org/10.1080/01621459.2016.1165102

HEPLER, A., SAUNDERS, C. P., DAVIS, L. and BUSCAGLIA, J. (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Sci. Int.* **219** 129–140.

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A, Math. Phys. Sci.* **186** 453–461. MR0017504 https://doi.org/10.1098/rspa.1946.0056

KAFADAR, K. (2018). The critical role of statistics in demonstrating the reliability of expert evidence. *Fordham Law Rev.* **86** 1617–1637.

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

LINDLEY, D. V. (1971). *Bayesian Statistics, a Review. Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No.* 2. SIAM, Philadelphia, PA. MR0329081

LINDLEY, D. V. (1977). A problem in forensic science. *Biometrika* **64** 207–213. MR0518935 https://doi.org/10.1093/biomet/64.2.207

LUCY, D. (2013). comparison: Multivariate likelihood ratio calculation and evaluation. R package version 1.0-4. https://CRAN.R-project.org/package=comparison.

LUND, S. P. and IYER, H. (2017). Likelihood ratio as weight of forensic evidence: A closer look. *J. Res. Natl. Inst. Stand. Technol.* **122** 1–32.

MARTIN, R. (2021). Inferential models and the decision-theoretic implications of the validity property. Preprint. Available at arXiv:2112.13247.

MARTIN, R. and LIU, C. (2016). *Inferential Models: Reasoning with Uncertainty. Monographs on Statistics and Applied Probability* **147**. CRC Press, Boca Raton, FL. MR3618727

MARTIN, R. and WALKER, S. G. (2019). Data-driven priors and their posterior concentration rates. *Electron. J. Stat.* **13** 3049–3081. MR4010592 https://doi.org/10.1214/19-ejs1600

MENG, X.-L. and WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860. MR1422406

MORRISON, G. S. (2016). Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate. *Sci. Justice* **56** 371–373.

MUKERJEE, R. and REID, N. (1999). On a property of probability matching priors: Matching the alternative coverage probabilities. *Biometrika* **86** 333–340. MR1705343 https://doi.org/10.1093/biomet/86.2.333

MURPH, A., HANNIG, J. and WILLIAMS, J. (2020). Introduction to Generalized Fiducial Inference. In *CRC Press BFF Handbook*. To appear.

NEUMANN, C. and AUSDEMORE, M. A. (2020). Defence against the modern arts: the curse of statistics—Part II: 'Score-based likelihood ratios.' *Law Probab. Risk* **19** 21–42. https://doi.org/10.1093/lpr/mgaa006

NEUMANN, C., CHAMPOD, C., PUCH-SOLIS, R., EGLI, N. M., ANTHONIOZ, A. and BROMAGE-GRIFFITHS, A. (2007). Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *J. Forensic Sci.* **52** 54–64.

NEUMANN, C., HENDRICKS, J. and AUSDEMORE, M. A. (2020). Statistical support for conclusions in fingerprint examinations. In *Handbook of Forensic Statistics*, CRC Press, Boca Raton, FL, USA.

OMMEN, D. M. (2017). Approximate statistical solutions to the forensic identification of source problem. Electronic Theses and Dissertations 1710.

OMMEN, D. M. and SAUNDERS, C. P. (2019). Reconciling the bayes factor and likelihood ratio for two non-nested model selection problems. Preprint. Available at arXiv:1901.09798.

OMMEN, D. M., SAUNDERS, C. P. and NEUMANN, C. (2017). The characterization of Monte Carlo errors for the quantification of the value of forensic evidence. *J. Stat. Comput. Simul.* **87** 1608–1643. MR3625235 https://doi.org/10.1080/00949655.2017.1280036

PARKER, J. B. (1966). A statistical treatment of identification problems. *J. - Forensic Sci. Soc.* **6** 33–39.

PINHEIRO, J., BATES, D., DEBROY, S. and SARKAR, D. (2019). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-140. https://CRAN.R-project.org/package=nlme.

RAMOS, D. and GONZALEZ-RODRIGUEZ, J. (2008). Cross-entropy analysis of the information in forensic speaker recognition. In *'Odyssey* 2008: *The Speaker and Language Recognition Workshop'*, *International Speech Communication Association*.

SAVAGE, L. J. (1961). The foundations of statistics reconsidered. In *Proc.* 4*th Berkeley Sympos. Math. Statist. and Prob.*, *Vol. I*: *Contributions to the Theory of Statistics* 575–586. Univ. California Press, Berkeley, CA. MR0133898

SHAFER, G. (1982). Lindley's paradox. *J. Amer. Statist. Assoc.* **77** 325–351. MR0664677

SHI, W. J., HANNIG, J., LAI, R. C. S. and LEE, T. C. M. (2021). Covariance estimation via fiducial inference. *Stat. Theory Relat. Fields* **5** 316–331. MR4335036 https://doi.org/10.1080/24754269.2021.1877950

STAICU, A.-M. and REID, N. M. (2008). On probability matching priors. *Canad. J. Statist.* **36** 613–622. MR2532255 https://doi.org/10.1002/cjs.5550360408

SWOFFORD, H., KOERNTNER, A., ZEMP, F., AUSDEMORE, M., LIU, A. and SALYARDS, M. (2018). A method for the statistical interpretation of friction ridge skin impression evidence: Method development and validation. *Forensic Sci. Int.* **287** 113–126.

TARALDSEN, G. and LINDQVIST, B. H. (2013). Fiducial theory and optimal inference. *Ann. Statist.* **41** 323–341. MR3059420 https://doi.org/10.1214/13-AOS1083

TARALDSEN, G. and LINDQVIST, B. (2021). Fiducial inference and decision theory. Preprint. Available at arXiv:2112.07060.

TARONI, F., BOZZA, S., BIEDERMANN, A. and AITKEN, C. G. G. (2016). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law Probab. Risk* **15** 1–16.

VAN ES, A., WIARDA, W., HORDIJK, M., ALBERINK, I. and VERGEER, P. (2017). Implementation and assessment of a likelihood ratio approach for the evaluation of la-icp-ms evidence in forensic glass analysis. *Sci. Justice* **57** 181–192.

WASSERSTEIN, R. L. and LAZAR, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose [Editorial]. *Amer. Statist.* **70** 129–133. MR3511040 https://doi.org/10.1080/00031305.2016.1154108

WILLIAMS, J. P. and HANNIG, J. (2019). Nonpenalized variable selection in high-dimensional linear model settings via generalized fiducial inference. *Ann. Statist.* **47** 1723–1753. MR3911128 https://doi.org/10.1214/18-AOS1733

WILLIAMS, J. P., OMMEN, D. M. and HANNIG, J. (2023). Supplement to "Generalized fiducial factor: an alternative to the Bayes factor for forensic identification of source problems." https://doi.org/10.1214/22-AOAS1632SUPP

WILLIAMS, J. P., XIE, Y. and HANNIG, J. (2019). The EAS approach for graphical selection consistency in vector autoregression models. *Canad. J. Statist.* 1–28. To appear.

ZADORA, G., MARTYNA, A., RAMOS, D. and AITKEN, C. (2013). *Statistical Analysis in Forensic Science*: *Evidential Value of Multivariate Physicochemical Data*. Wiley, New York.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 403–433
https://doi.org/10.1214/22-AOAS1633

TOPOLOGICAL LEARNING FOR BRAIN NETWORKS

BY TANANUN SONGDECHAKRAIWUT[a] AND MOO K. CHUNG[b]

Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, [a]*songdechakra@wisc.edu,* [b]*mkchung@wisc.edu*

This paper proposes a novel topological learning framework that integrates networks of different sizes and topology through persistent homology. Such challenging task is made possible through the introduction of a computationally efficient topological loss. The use of the proposed loss bypasses the intrinsic computational bottleneck associated with matching networks. We validate the method in extensive statistical simulations to assess its effectiveness when discriminating networks with different topology. The method is further demonstrated in a twin brain imaging study where we determine if brain networks are genetically heritable. The challenge here is due to the difficulty of overlaying the topologically different functional brain networks obtained from resting-state functional MRI onto the template structural brain network obtained through diffusion MRI.

REFERENCES

ADLER, R. J., AGAMI, S. and PRANAV, P. (2017). Modeling and replicating statistical topology and evidence for CMB nonhomogeneity. *Proc. Natl. Acad. Sci. USA* **114** 11878–11883. MR3725115 https://doi.org/10.1073/pnas.1706885114

ADLER, R. J., BOBROWSKI, O., BORMAN, M. S., SUBAG, E. and WEINBERGER, S. (2010). Persistent homology for random fields and complexes. In *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown. Inst. Math. Stat. (IMS) Collect.* **6** 124–143. IMS, Beachwood, OH. MR2798515

ANDERSSON, J. L., SKARE, S. and ASHBURNER, J. (2003). How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging. *NeuroImage* **20** 870–888.

ARSLAN, S., KTENA, S. I., MAKROPOULOS, A., ROBINSON, E. C., RUECKERT, D. and PARISOT, S. (2018). Human brain mapping: A systematic comparison of parcellation methods for the human cerebral cortex. *NeuroImage* **170** 5–30. https://doi.org/10.1016/j.neuroimage.2017.04.014

AVANTS, B. B., EPSTEIN, C. L., GROSSMAN, M. and GEE, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12** 26–41. https://doi.org/10.1016/j.media.2007.06.004

AVANTS, B. B., TUSTISON, N. J., SONG, G., COOK, P. A., KLEIN, A. and GEE, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage* **54** 2033–2044.

BABAI, L. and LUKS, E. M. (1983). Canonical labeling of graphs. In *Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing* 171–183.

BENDICH, P., MARRON, J. S., MILLER, E., PIELOCH, A. and SKWERER, S. (2016). Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.* **10** 198–218. MR3480493 https://doi.org/10.1214/15-AOAS886

BLOKLAND, G. A. M., MCMAHON, K. L., THOMPSON, P. M., MARTIN, N. G., DE ZUBICARAY, G. I. and WRIGHT, M. J. (2011). Heritability of working memory brain activation. *J. Neurosci.* **31** 10882–10890.

BONNEEL, N., RABIN, J., PEYRÉ, G. and PFISTER, H. (2015). Sliced and Radon Wasserstein barycenters of measures. *J. Math. Imaging Vision* **51** 22–45. MR3300482 https://doi.org/10.1007/s10851-014-0506-3

BRYANT, C., ZHU, H., AHN, M. and IBRAHIM, J. (2017). LCN: A random graph mixture model for community detection in functional brain networks. *Stat. Interface* **10** 369–378. MR3608546 https://doi.org/10.4310/SII.2017.v10.n3.a1

BULLMORE, E. and SPORNS, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10** 186–198.

CABALLERO-GAUDES, C. and REYNOLDS, R. C. (2017). Methods for cleaning the BOLD fMRI signal. *NeuroImage* **154** 128–149. https://doi.org/10.1016/j.neuroimage.2016.12.018

Key words and phrases. Topological data analysis, persistent homology, topological learning, Wasserstein distance, birth-death decomposition, twin brain imaging study.

CALLAGHAN, P. T., ECCLES, C. D. and XIA, Y. (1988). NMR microscopy of dynamic displacements: k-space and q-space imaging. *J. Phys. E, Sci. Instrum.* **21** 820.

CARLSSON, G. (2009). Topology and data. *Bull. Amer. Math. Soc.* (*N.S.*) **46** 255–308. MR2476414 https://doi.org/10.1090/S0273-0979-09-01249-X

CARLSSON, G. and MÉMOLI, F. (2010). Characterization, stability and convergence of hierarchical clustering methods. *J. Mach. Learn. Res.* **11** 1425–1470. MR2645457

CARRIERE, M., CUTURI, M. and OUDOT, S. (2017). Sliced Wasserstein kernel for persistence diagrams. In *International Conference on Machine Learning* 664–673. PMLR.

CHAZAL, F., COHEN-STEINER, D., GUIBAS, L. J., MÉMOLI, F. and OUDOT, S. Y. (2009). Gromov–Hausdorff stable signatures for shapes using persistence. In *Computer Graphics Forum* **28** 1393–1403. Wiley Online Library.

CHAZAL, F., GUIBAS, L. J., OUDOT, S. Y. and SKRABA, P. (2013). Persistence-based clustering in Riemannian manifolds. *J. ACM* **60** Art. 41, 38 pp. MR3144911 https://doi.org/10.1145/2535927

CHEN, C., NI, X., BAI, Q. and WANG, Y. (2019). A topological regularizer for classifiers via persistent homology. In *The 22nd International Conference on Artificial Intelligence and Statistics* 2573–2582.

CHIANG, M. C., MCMAHON, K. L., DE ZUBICARAY, G. I., MARTIN, N. G., HICKIE, I., TOGA, A. W., WRIGHT, M. J. and THOMPSON, P. M. (2011). Genetics of white matter development: A DTI study of 705 twins and their siblings aged 12 to 29. *NeuroImage* **54** 2308–2317.

CHO, M., LEE, J. and LEE, K. M. (2010). Reweighted random walks for graph matching. In *European Conference on Computer Vision* 492–505. Springer, Berlin.

CHRISTIAENS, D., REISERT, M., DHOLLANDER, T., SUNAERT, S., SUETENS, P. and MAES, F. (2015). Global tractography of multi-shell diffusion-weighted imaging data using a multi-tissue model. *NeuroImage* **123** 89–101. https://doi.org/10.1016/j.neuroimage.2015.08.008

CHUNG, M. K. and OMBAO, H. (2021). Lattice paths for persistent diagrams. In *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data*. *LNCS* **12929** 77–86.

CHUNG, M. K., ADLURU, N., DALTON, K. M., ALEXANDER, A. L. and DAVIDSON, R. J. (2011). Scalable brain network construction on white matter fibers. *Proc. SPIE Int. Soc. Opt. Eng.* **7962** 1403–1408. https://doi.org/10.1117/12.874245

CHUNG, M. K., HUANG, S.-G., GRITSENKO, A., SHEN, L. and LEE, H. (2019a). Statistical inference on the number of cycles in brain networks. In 2019 *IEEE 16th International Symposium on Biomedical Imaging* (*ISBI* 2019) 113–116. IEEE, New York.

CHUNG, M. K., LEE, H., DICHRISTOFANO, A., OMBAO, H. and SOLO, V. (2019b). Exact topological inference of the resting-state brain networks in twins. *Netw. Neurosci.* **3** 674–694.

CHUNG, M. K., XIE, L., HUANG, S.-G., WANG, Y., YAN, J. and SHEN, L. (2019c). Rapid acceleration of the permutation test via transpositions. In *International Workshop on Connectomics in NeuroImaging* 42–53. Springer, Berlin.

CHUNG, M. K., LEE, H. SOLO, V. DAVIDSON, R. J. POLLAK, S. D. (2017a). Topological distances between brain networks. In *Connectomics in Neuroimaging*: *First International Workshop* **10511** 161–170. Springer, Berlin. https://doi.org/10.1007/978-3-319-67159-8_19

CHUNG, M. K., VILLALTA-GIL, V., LEE, H., RATHOUZ, P. J., LAHEY, B. B. and ZALD, D. H. (2017b). Exact topological inference for paired brain networks via persistent homology. In *International Conference on Information Processing in Medical Imaging* 299–310. Springer, Berlin.

CLOUGH, J. R., OKSUZ, I., BYRNE, N., SCHNABEL, J. A. and KING, A. P. (2019). Explicit topological priors for deep-learning based image segmentation using persistent homology. In *International Conference on Information Processing in Medical Imaging* 16–28. Springer, Berlin.

COHEN-STEINER, D., EDELSBRUNNER, H. and HARER, J. (2007). Stability of persistence diagrams. *Discrete Comput. Geom.* **37** 103–120. MR2279866 https://doi.org/10.1007/s00454-006-1276-5

COHEN-STEINER, D., EDELSBRUNNER, H., HARER, J. and MILEYKO, Y. (2010). Lipschitz functions have L_p-stable persistence. *Found. Comput. Math.* **10** 127–139. MR2594441 https://doi.org/10.1007/s10208-010-9060-6

COUR, T., SRINIVASAN, P. and SHI, J. (2006). Balanced graph matching. *Adv. Neural Inf. Process. Syst.* **19** 313–320.

CRAWFORD, L., MONOD, A., CHEN, A. X., MUKHERJEE, S. and RABADÁN, R. (2020). Predicting clinical outcomes in glioblastoma: An application of topological and functional data analysis. *J. Amer. Statist. Assoc.* **115** 1139–1150. MR4143455 https://doi.org/10.1080/01621459.2019.1671198

DESHPANDE, I., ZHANG, Z. and SCHWING, A. G. (2018). Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3483–3491.

DESIKAN, R. S., SÉGONNE, F., FISCHL, B., QUINN, B. T., DICKERSON, B. C., BLACKER, D., BUCKNER, R. L., DALE, A. M., MAGUIRE, R. P. et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31** 968–980.

DIEDRICHSEN, J. and SHADMEHR, R. (2005). Detecting and adjusting for artifacts in fMRI time series data. *NeuroImage* **27** 624–634.

EDELSBRUNNER, H. and HARER, J. (2008). Persistent homology—A survey. In *Surveys on Discrete and Computational Geometry*. *Contemp*. *Math*. **453** 257–282. Amer. Math. Soc., Providence, RI. MR2405684 https://doi.org/10.1090/conm/453/08802

EDELSBRUNNER, H., LETSCHER, D. and ZOMORODIAN, A. (2000). Topological persistence and simplification. In 41*st Annual Symposium on Foundations of Computer Science* (*Redondo Beach*, *CA*, 2000) 454–463. IEEE Comput. Soc. Press, Los Alamitos, CA. MR1931842 https://doi.org/10.1109/SFCS.2000.892133

EDMONDS, J. and KARP, R. M. (1972). Theoretical improvements in algorithmic efficiency for network flow problems. *J. ACM* **19** 248–264.

EICKHOFF, S. B., NICHOLS, T. E., LAIRD, A. R., HOFFSTAEDTER, F., AMUNTS, K., FOX, P. T., BZDOK, D. and EICKHOFF, C. R. (2016). Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. *NeuroImage* **137** 70–85. https://doi.org/10.1016/j.neuroimage.2016.04.072

EICKHOFF, S. B., YEO, B. T. T. and GENON, S. (2018). Imaging-based parcellations of the human brain. *Nat*. *Rev*. *Neurosci*. **19** 672–686.

FALCONER, D. and MACKAY, T. (1995). *Introduction to Quantitative Genetics*, 4th ed. Longman, Harlow.

FAN, L., LI, H., ZHUO, J., ZHANG, Y., WANG, J., CHEN, L., YANG, Z., CHU, C., XIE, S. et al. (2016). The human brainnetome atlas: A new brain atlas based on connectional architecture. *Cereb. Cortex* **26** 3508–3526.

FORNITO, A., ZALESKY, A. and BULLMORE, E. T. (2010). Network scaling effects in graph analytic studies of human resting-state fMRI data. *Front*. *Syst*. *Neurosci*. **4** 1–16.

FORNITO, A., ZALESKY, A. and BULLMORE, E. (2016). *Fundamentals of Brain Network Analysis*. Academic Press, New York.

GHRIST, R. (2008). Barcodes: The persistent topology of data. *Bull*. *Amer*. *Math*. *Soc*. (*N.S.*) **45** 61–75. MR2358377 https://doi.org/10.1090/S0273-0979-07-01191-3

GINESTET, C. E., NICHOLS, T. E., BULLMORE, E. T. and SIMMONS, A. (2011). Brain network analysis: Separating cost from topology using cost-integration. *PLoS ONE* **6** e21570. https://doi.org/10.1371/journal.pone.0021570

GLAHN, D. C., WINKLER, A. M., KOCHUNOV, P., ALMASY, L., DUGGIRALA, R., CARLESS, M. A., CURRAN, J. C., OLVERA, R. L., LAIRD, A. R. et al. (2010). Genetic control over the resting brain. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **107** 1223–1228.

GLASSER, M. F. and ESSEN, D. C. V. (2011). Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI. *J. Neurosci*. **31** 11597–11616. https://doi.org/10.1523/JNEUROSCI.2180-11.2011

GLASSER, M. F., SMITH, S. M., MARCUS, D. S., ANDERSSON, J. L. R., AUERBACH, E. J., BEHRENS, T. E. J., COALSON, T. S., HARMS, M. P., JENKINSON, M. et al. (2016). The human connectome project's neuroimaging approach. *Nat*. *Neurosci*. **19** 1175.

GLASSER, M. F., SOTIROPOULOS, S. N., WILSON, J. A., COALSON, T. S., FISCHL, B., ANDERSSON, J. L., XU, J., JBABDI, S., WEBSTER, M. et al. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80** 105–124.

GOLD, S. and RANGARAJAN, A. (1996). A graduated assignment algorithm for graph matching. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **18** 377–388.

GONG, G., HE, Y., CONCHA, L., LEBEL, C., GROSS, D. W., EVANS, A. C. and BEAULIEU, C. (2009). Mapping anatomical connectivity patterns of human cerebral cortex using in vivo diffusion tensor imaging tractography. *Cereb. Cortex* **19** 524–536.

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3617773

GRITSENKO, A., LINDQUIST, M. A. and CHUNG, M. K. (2020). Twin classification in resting-state brain connectivity. In *IEEE International Symposium on Biomedical Imaging* (*ISBI*). Available at arXiv:1807.00244.

GUO, X. and SRIVASTAVA, A. (2020). Representations, metrics and statistics for shape analysis of elastic graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 832–833.

HAGMANN, P., KURANT, M., GIGANDET, X., THIRAN, P., WEDEEN, V. J., MEULI, R. and THIRAN, J.-P. (2007). Mapping human whole-brain structural networks with diffusion MRI. *PLoS ONE* **2** e597. https://doi.org/10.1371/journal.pone.0000597

HONEY, C. J., KÖTTER, R., BREAKSPEAR, M. and SPORNS, O. (2007). Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **104** 10240–10245.

HONEY, C. J., SPORNS, O., CAMMOUN, L., GIGANDET, X., THIRAN, J.-P., MEULI, R. and HAGMANN, P. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proc. Natl. Acad. Sci. USA* **106** 2035–2040.

HU, X., LI, F., SAMARAS, D. and CHEN, C. (2019). Topology-preserving deep image segmentation. In *Advances in Neural Information Processing Systems* 5657–5668.

HUANG, S. G., CHUNG, M. K., QIU, A. and INITIATIVE, A. D. N. (2021). Revisiting convolutional neural network on graphs with polynomial approximations of Laplace–Beltrami spectral filtering. *Neural Comput. Appl.* **33** 13693–13704.

HUANG, S. G., SAMDIN, S. T., TING, C. M., OMBAO, H. and CHUNG, M. K. (2020). Statistical model for dynamically-changing correlation matrices with application to brain connectivity. *J. Neurosci. Methods* **331** 108480.

JENKINSON, M. and SMITH, S. (2001). A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* **5** 143–156.

JENKINSON, M., BANNISTER, P., BRADY, M. and SMITH, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* **17** 825–841.

JEURISSEN, B., TOURNIER, J.-D., DHOLLANDER, T., CONNELLY, A. and SIJBERS, J. (2014). Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data. *NeuroImage* **103** 411–426. https://doi.org/10.1016/j.neuroimage.2014.07.061

JOVICICH, J., CZANNER, S., GREVE, D., HALEY, E., VAN DER KOUWE, A., GOLLUB, R., KENNEDY, D., SCHMITT, F., BROWN, G. et al. (2006). Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *NeuroImage* **30** 436–443.

KANG, H., OMBAO, H., FONNESBECK, C., DING, Z. and MORGAN, V. L. (2017). A Bayesian double fusion model for resting-state brain connectivity using joint functional and structural data. *Brain Connect.* **7** 219–227. https://doi.org/10.1089/brain.2016.0447

KARAS, M., BRZYSKI, D., DZEMIDZIC, M., GOÑI, J., KAREKEN, D. A., RANDOLPH, T. W. and HAREZLAK, J. (2019). Brain connectivity-informed regularization methods for regression. *Stat. Biosci.* **11** 47–90. https://doi.org/10.1007/s12561-017-9208-x

KARRAS, T., AILA, T., LAINE, S. and LEHTINEN, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*.

KERBER, M., MOROZOV, D. and NIGMETOV, A. (2017). Geometry helps to compare persistence diagrams. *ACM J. Exp. Algorithmics* **22** Art. 1.4, 20 pp. MR3707737 https://doi.org/10.1145/3064175

KOLOURI, S., NADJAHI, K., SIMSEKLI, U., BADEAU, R. and ROHDE, G. (2019). Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems* 261–272.

KOLOURI, S., PARK, S. R., THORPE, M., SLEPCEV, D. and ROHDE, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Process. Mag.* **34** 43–59.

KONG, R., GAO, J., XU, Y., PAN, Y., WANG, J. and LIU, J. (2019). Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. *Neurocomputing* **324** 63–68.

KRIEGEL, H.-P., KRÖGER, P. and ZIMEK, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data* **3** 1–58.

LE, H. and KUME, A. (2000). The Fréchet mean shape and the shape of the means. *Adv. in Appl. Probab.* **32** 101–113. MR1765168 https://doi.org/10.1239/aap/1013540025

LEE, H., KANG, H., CHUNG, M. K., KIM, B.-N. and LEE, D. S. (2012). Persistent brain network homology from the perspective of dendrogram. *IEEE Trans. Med. Imag.* **31** 2267–2277.

LEORDEANU, M. and HEBERT, M. (2005). A spectral technique for correspondence problems using pairwise constraints. In *Tenth IEEE International Conference on Computer Vision* (*ICCV'*05), *Volume* 1 **2** 1482–1489. IEEE, New York.

LEORDEANU, M., HEBERT, M. and SUKTHANKAR, R. (2009). An integer projected fixed point method for graph matching and map inference. In *Advances in Neural Information Processing Systems* 1114–1122.

LIUTKUS, A., SIMSEKLI, U., MAJEWSKI, S., DURMUS, A. and STÖTER, F.-R. (2019). Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning* 4104–4113.

LOVE, E. R., FILIPPENKO, B., MAROULAS, V. and CARLSSON, G. (2021). Topological deep learning. Preprint. Available at arXiv:2101.05778.

LV, J., GUO, L., HU, X., ZHANG, T., LI, K., ZHANG, D., YANG, J. and LIU, T. (2010). Fiber-centered analysis of brain connectivities using DTI and resting state FMRI data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (*MICCAI*) 143–150. Springer, Berlin.

MARCHESE, A. and MAROULAS, V. (2018). Signal classification with a point process distance on the space of persistence diagrams. *Adv. Data Anal. Classif.* **12** 657–682. MR3854422 https://doi.org/10.1007/s11634-017-0294-x

MARCOS, D., VOLPI, M. and TUIA, D. (2016). Learning rotation invariant convolutional filters for texture classification. In 2016 23*rd International Conference on Pattern Recognition* (*ICPR*) 2012–2017. IEEE, New York.

MAROULAS, V., NASRIN, F. and OBALLE, C. (2020). A Bayesian framework for persistent homology. *SIAM J. Math. Data Sci.* **2** 48–74. MR4060450 https://doi.org/10.1137/19M1268719

MCKAY, D. R., KNOWLES, E. E. M., WINKLER, A. A. M., SPROOTEN, E., KOCHUNOV, P., OLVERA, R. L., CURRAN, J. E., KENT JR., J. W., CARLESS, M. A. et al. (2014). Influence of age, sex and genetic factors on the human brain. *Brain Imaging Behav.* **8** 143–152.

NAITZAT, G., ZHITNIKOV, A. and LIM, L.-H. (2020). Topology of deep neural networks. *J. Mach. Learn. Res.* **21** Paper No. 184, 40 pp. MR4209470

NG, A. Y., JORDAN, M. I. and WEISS, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* 849–856.

OMBAO, H., LINDQUIST, M., THOMPSON, W. and ASTON, J., eds. (2016). *Handbook of Neuroimaging Data Analysis*. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR3618824

PATRANGENARU, V., BUBENIK, P., PAIGE, R. L. and OSBORNE, D. (2019). Challenges in topological object data analysis. *Sankhya A* **81** 244–271. MR3982197 https://doi.org/10.1007/s13171-018-0137-7

POWER, J. D., BARNES, K. A., SNYDER, A. Z., SCHLAGGAR, B. L. and PETERSEN, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage* **59** 2142–2154.

RABIN, J., PEYRÉ, G., DELON, J. and BERNOT, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision* 435–446. Springer, Berlin.

REININGHAUS, J., HUBER, S., BAUER, U. and KWITT, R. (2015). A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4741–4748.

REYNOLDS, C. A. and PHILLIPS, D. (2015). Genetics of brain aging–twin aging.

ROBINS, V. and TURNER, K. (2016). Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids. *Phys. D* **334** 99–117. MR3545972 https://doi.org/10.1016/j.physd.2016.03.007

ROTTSCHY, C., LANGNER, R., DOGAN, I., REETZ, K., LAIRD, A. R., SCHULZ, J. B., FOX, P. T. and EICKHOFF, S. B. (2012). Modelling neural correlates of working memory: A coordinate-based meta-analysis. *NeuroImage* **60** 830–846.

SAAD, D., ed. (1998). *On-Line Learning in Neural Networks*. *Publications of the Newton Institute* **17**. Cambridge Univ. Press, Cambridge. MR1763695

SATTERTHWAITE, T. D., WOLF, D. H., LOUGHEAD, J., RUPAREL, K., ELLIOTT, M. A., HAKONARSON, H., GUR, R. C. and GUR, R. E. (2012). Impact of in-scanner head motion on multiple measures of functional connectivity: Relevance for studies of neurodevelopment in youth. *NeuroImage* **60** 623–632.

SCHAEFER, A., KONG, R., GORDON, E. M., LAUMANN, T. O., ZUO, X. N., HOLMES, A. J., EICKHOFF, S. B. and YEO, B. T. T. (2017). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* **28** 3095–3114.

SCOTT, J. (1988). Social network analysis. *Sociology* **22** 109–127.

SHAH, L. M., CRAMER, J. A., FERGUSON, M. A., BIRN, R. M. and ANDERSON, J. S. (2016). Reliability and reproducibility of individual differences in functional connectivity acquired during task and resting state. *Brain Behav.* **6** e00456.

SHAPPELL, H. M., CAFFO, B. S., PEKAR, J. J. and LINDQUIST, M. (2019). Improved state change estimation in dynamic functional connectivity using hidden semi-Markov models. *NeuroImage* **191** 243–257.

SHARATHKUMAR, R. and AGARWAL, P. K. (2012). Algorithms for the transportation problem in geometric settings. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms* 306–317. ACM, New York. MR3205219

SHATTUCK, D. W., MIRZA, M., ADISETIYO, V., HOJATKASHANI, C., SALAMON, G., NARR, K. L., POLDRACK, R. A., BILDER, R. M. and TOGA, A. W. (2008). Construction of a 3D probabilistic atlas of human cortical structures. *NeuroImage* **39** 1064–1080.

SINGH, N., COUTURE, H. D., MARRON, J., PEROU, C. and NIETHAMMER, M. (2014). Topological descriptors of histology images. In *International Workshop on Machine Learning in Medical Imaging* 231–239. Springer, Berlin.

SMIT, D. J. A., STAM, C. J., POSTHUMA, D., BOOMSMA, D. I. and DE GEUS, E. J. C. (2008). Heritability of small-world networks in the brain: A graph theoretical analysis of resting-state EEG functional connectivity. *Hum. Brain Mapp.* **29** 1368–1378.

SMITH, R. E., TOURNIER, J. D., CALAMANTE, F. and CONNELLY, A. (2015). SIFT2: Enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography. *NeuroImage* **119** 338–351.

SNIJDERS, T., SPREEN, M. and ZWAAGSTRA, R. (1995). The use of multilevel modeling for analysing personal networks: Networks of cocaine users in an urban area. *J. Quant. Anthropol.* **5** 85–105.

SONGDECHAKRAIWUT, T. and CHUNG, M. K. (2020). Dynamic topological data analysis for functional brain signals. In 2020 *IEEE* 17*th International Symposium on Biomedical Imaging Workshops* (*ISBI Workshops*) 1–4. IEEE, New York.

SONGDECHAKRAIWUT, T., KRAUSE, B. M., BANKS, M. I., NOURSKI, K. V. and VAN VEEN, B. D. (2022a). Topological classification in a Wasserstein distance based vector space. Preprint. Available at arXiv:2202.01275.

SONGDECHAKRAIWUT, T., KRAUSE, B. M., BANKS, M. I., NOURSKI, K. V. and VEEN, B. D. V. (2022b). Fast topological clustering with Wasserstein distance. In *International Conference on Learning Representations*.

SONGDECHAKRAIWUT, T., SHEN, L. and CHUNG, M. K. (2021). Topological learning and its application to multimodal brain network integration. In *Medical Image Computing and Computer Assisted Intervention* (*MICCAI*). *Lecture Notes in Computer Science* **12902** 166–176.

SPORNS, O. (2003). Graph Theory Methods for the Analysis of Neural Connectivity Patterns. In *Neuroscience Databases*: *A Practical Guide* 171–185. Springer, Boston, MA.

TAYLOR, L. and NITSCHKE, G. (2018). Improving deep learning with generic data augmentation. In 2018 *IEEE Symposium Series on Computational Intelligence* (*SSCI*) 1542–1547. IEEE, New York.

TIAN, Y., YAN, J., ZHANG, H., ZHANG, Y., YANG, X. and ZHA, H. (2012). On the convergence of graph matching: Graduated assignment revisited. In *European Conference on Computer Vision* 821–835. Springer, Berlin.

TOGA, A. W. and THOMPSON, P. M. (2003). Mapping brain asymmetry. *Nat. Rev. Neurosci.* **4** 37–48. https://doi.org/10.1038/nrn1009

TOURNIER, J., CALAMANTE, F., CONNELLY, A. et al. (2012). MRtrix: Diffusion tractography in crossing fiber regions. *Int. J. Imaging Syst. Technol.* **22** 53–66.

TURNER, K., MILEYKO, Y., MUKHERJEE, S. and HARER, J. (2014). Fréchet means for distributions of persistence diagrams. *Discrete Comput. Geom.* **52** 44–70. MR3231030 https://doi.org/10.1007/s00454-014-9604-7

TZOURIO-MAZOYER, N., LANDEAU, B., PAPATHANASSIOU, D., CRIVELLO, F., ETARD, O., DELCROIX, N., MAZOYER, B. and JOLIOT, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15** 273–289.

VAN DIJK, K. R. A., SABUNCU, M. R. and BUCKNER, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *NeuroImage* **59** 431–438.

VAN ESSEN, D. C., UGURBIL, K., AUERBACH, E., BARCH, D., BEHRENS, T., BUCHOLZ, R., CHANG, A., CHEN, L., CORBETTA, M. et al. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage* **62** 2222–2231.

VAN ESSEN, D. C., SMITH, S. M., BARCH, D. M., BEHRENS, T. E., YACOUB, E., UGURBIL, K., CONSORTIUM, W.-M. H. et al. (2013). The WU-Minn human connectome project: An overview. *NeuroImage* **80** 62–79.

WANG, Y., OMBAO, H. and CHUNG, M. K. (2018). Topological data analysis of single-trial electroencephalographic signals. *Ann. Appl. Stat.* **12** 1506–1534. MR3852686 https://doi.org/10.1214/17-AOAS1119

WANG, T., LIU, H., LI, Y., JIN, Y., HOU, X. and LING, H. (2020). Learning combinatorial solver for graph matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7568–7577.

WASSERMAN, L. (2018). Topological data analysis. *Annu. Rev. Stat. Appl.* **5** 501–535. MR3774757 https://doi.org/10.1146/annurev-statistics-031017-100045

WIJK, B. C. M., STAM, C. J. and DAFFERTSHOFER, A. (2010). Comparing brain networks of different size and connectivity density using graph theory. *PLoS ONE* **5** e13701.

XIA, K. and WEI, G.-W. (2014). Persistent homology analysis of protein structure, flexibility, and folding. *Int. J. Numer. Methods Biomed. Eng.* **30** 814–844. MR3247713 https://doi.org/10.1002/cnm.2655

XIE, L., AMICO, E., SALAMA, P., WU, Y. C., FANG, S., SPORNS, O., SAYKIN, A. J., GOÑI, J., YAN, J. et al. (2018). Heritability estimation of reliable connectomic features. In *International Workshop on Connectomics in Neuroimaging*. *Lecture Notes in Computer Science* **11083** 58–66.

XUE, W., BOWMAN, F. D., PILEGGI, A. V. and MAYER, A. R. (2015). A multimodal approach for determining brain networks by jointly modeling functional and structural connectivity. *Front. Comput. Neurosci.* **9** 22. https://doi.org/10.3389/fncom.2015.00022

YU, T., YAN, J., WANG, Y., LIU, W. et al. (2018). Generalizing graph matching beyond quadratic assignment model. In *Advances in Neural Information Processing Systems* 853–863.

ZALESKY, A., FORNITO, A., HARDING, I. H., COCCHI, L., YÜCEL, M., PANTELIS, C. and BULLMORE, E. T. (2010). Whole-brain anatomical networks: Does the choice of nodes matter? *NeuroImage* **50** 970–983.

ZAVLANOS, M. M. and PAPPAS, G. J. (2008). A dynamical systems approach to weighted graph matching. *Automatica J. IFAC* **44** 2817–2824. MR2527202 https://doi.org/10.1016/j.automatica.2008.04.009

ZEMEL, Y. and PANARETOS, V. M. (2019). Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli* **25** 932–976. MR3920362 https://doi.org/10.3150/17-bej1009

ZHANG, Z., DESCOTEAUX, M., ZHANG, J., GIRARD, G., CHAMBERLAND, M., DUNSON, D., SRIVASTAVA, A. and ZHU, H. (2018). Mapping population-based structural connectomes. *NeuroImage* **172** 130–145.

ZHANG, G., CAI, B., ZHANG, A., STEPHEN, J. M., WILSON, T. W., CALHOUN, V. D. and WANG, Y. P. (2019a). Estimating dynamic functional brain connectivity with a sparse hidden Markov model. *IEEE Trans. Med. Imag.* **39** 488–498.

ZHANG, Z., XIANG, Y., WU, L., XUE, B. and NEHORAI, A. (2019b). KerGM: Kernelized graph matching. In *Advances in Neural Information Processing Systems* 3335–3346.

ZHOU, F. and DE LA TORRE, F. (2013). Deformable graph matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2922–2929.

ZHU, D., ZHANG, T., JIANG, X., HU, X., CHEN, H., YANG, N., LV, J., HAN, J., GUO, L. et al. (2014). Fusing DTI and fMRI data: A survey of methods and applications. *NeuroImage* **102** 184–191.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 434–453
https://doi.org/10.1214/22-AOAS1634

INDIVIDUALIZED RISK ASSESSMENT OF PREOPERATIVE OPIOID USE BY INTERPRETABLE NEURAL NETWORK REGRESSION

BY YUMING SUN[1,a], JIAN KANG[1,b], CHAD BRUMMETT[2,d] AND YI LI[1,c]

[1]*Department of Biostatistics, University of Michigan,* [a]*yumsun@umich.edu,* [b]*jiankang@umich.edu,* [c]*yili@umich.edu*
[2]*Department of Anesthesiology, University of Michigan,* [d]*cbrummet@med.umich.edu*

Preoperative opioid use has been reported to be associated with higher preoperative opioid demand, worse postoperative outcomes, and increased postoperative healthcare utilization and expenditures. Understanding the risk of preoperative opioid use helps establish patient-centered pain management. In the field of machine learning, deep neural network (DNN) has emerged as a powerful means for risk assessment because of its superb prediction power; however, the blackbox algorithms may make the results less interpretable than statistical models. Bridging the gap between the statistical and machine learning fields, we propose a novel interpretable neural network regression (INNER) which combines the strengths of statistical and DNN models. We use the proposed INNER to conduct individualized risk assessment of preoperative opioid use. Intensive simulations and an analysis of 34,186 patients expecting surgery in the Analgesic Outcomes Study (AOS) show that the proposed INNER not only can accurately predict the preoperative opioid use using preoperative characteristics as DNN but also can estimate the patient-specific odds of opioid use without pain and the odds ratio of opioid use for a unit increase in the reported overall body pain, leading to more straightforward interpretations of the tendency to use opioids than DNN. Our results identify the patient characteristics that are strongly associated with opioid use and is largely consistent with the previous findings, providing evidence that INNER is a useful tool for individualized risk assessment of preoperative opioid use.

REFERENCES

ARMAGHANI, S. J., LEE, D. S., BIBLE, J. E., ARCHER, K. R., SHAU, D. N., KAY, H., ZHANG, C., MCGIRT, M. J. and DEVIN, C. J. (2014). Preoperative opioid use and its association with perioperative opioid demand and postoperative opioid independence in patients undergoing spine surgery. *Spine* **39** E1524–E1530.

BAUER, B. and KOHLER, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.* **47** 2261–2285. MR3953451 https://doi.org/10.1214/18-AOS1747

BOTTOU, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'*2010 177–186. Physica-Verlag/Springer, Heidelberg. MR3362066

BROWN, L. A., DENIS, C. M., LEON, A., BLANK, M. B., DOUGLAS, S. D., MORALES, K. H., CRITS-CHRISTOPH, P. F., METZGER, D. S. and EVANS, D. L. (2021). Number of opioid overdoses and depression as a predictor of suicidal thoughts. *Drug Alcohol Depend.* **224** 108728.

BRUMMETT, C. M., JANDA, A. M., SCHUELLER, C. M., TSODIKOV, A., MORRIS, M., WILLIAMS, D. A. and CLAUW, D. J. (2013). Survey criteria for fibromyalgia independently predict increased postoperative opioid consumption after lower extremity joint arthroplasty: A prospective, observational cohort study. *Anesthesiology* **119** 1434–1443.

BRUMMETT, C. M., URQUHART, A. G., HASSETT, A. L., TSODIKOV, A., HALLSTROM, B. R., WOOD, N. I., WILLIAMS, D. A. and CLAUW, D. J. (2015). Characteristics of fibromyalgia independently predict poorer long-term analgesic outcomes following total knee and hip arthroplasty. *Arthritis & Rheumatology* **67** 1386–1394.

BRUMMETT, C. M., WALJEE, J. F., GOESLING, J., MOSER, S., LIN, P., ENGLESBE, M. J., BOHNERT, A. S., KHETERPAL, S. and NALLAMOTHU, B. K. (2017). New persistent opioid use after minor and major surgical procedures in US adults. *JAMA Surg.* **152** e170504(1)–e170504(9).

Key words and phrases. Deep learning, pain research, precision medicine, generalized linear models.

CAI, Z., FAN, J. and LI, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95** 888–902. MR1804446 https://doi.org/10.2307/2669472

CHE, Z., KALE, D., LI, W., BAHADORI, M. T. and LIU, Y. (2015). Deep computational phenotyping. In *Proceedings of the* 21*st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 507–516. ACM, New York.

CHE, Z., SAUVER, J. S., LIU, H. and LIU, Y. (2017). Deep learning solutions for classifying patients on opioid use. In *AMIA Annual Symposium Proceedings* **2017** 525–534. American Medical Informatics Association.

CHOI, E., BAHADORI, M. T., SCHUETZ, A., STEWART, W. F. and SUN, J. (2016). Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference* 301–318.

CRON, D. C., ENGLESBE, M. J., BOLTON, C. J., JOSEPH, M. T., CARRIER, K. L., MOSER, S. E., WALJEE, J. F., HILLIARD, P. E., KHETERPAL, S. et al. (2017). Preoperative opioid use is independently associated with increased costs and worse outcomes after major abdominal surgery. *Ann. Surg.* **265** 695–701.

DARKEN, C., CHANG, J. and MOODY, J. (1992). Learning rate schedules for faster stochastic gradient search. In *Neural Networks for Signal Processing II Proceedings of the* 1992 *IEEE Workshop* 3–12.

DONG, X., RASHIDIAN, S., WANG, Y., HAJAGOS, J., ZHAO, X., ROSENTHAL, R. N., KONG, J., SALTZ, M., SALTZ, J. et al. (2019). Machine learning based opioid overdose prediction using electronic health records. In *AMIA Annual Symposium Proceedings* **2019** 389–398. American Medical Informatics Association.

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. MR2054289 https://doi.org/10.1198/016214504000000089

EFRON, B. (2007a). Size, power and false discovery rates. *Ann. Statist.* **35** 1351–1377. MR2351089 https://doi.org/10.1214/009053606000001460

EFRON, B. (2007b). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302 https://doi.org/10.1198/016214506000001211

FAN, J., MA, C. and ZHONG, Y. (2021). A selective overview of deep learning. *Statist. Sci.* **36** 264–290. MR4255194 https://doi.org/10.1214/20-sts783

GOESLING, J., HENRY, M. J., MOSER, S. E., RASTOGI, M., HASSETT, A. L., CLAUW, D. J. and BRUMMETT, C. M. (2015). Symptoms of depression are associated with opioid use regardless of pain severity and physical functioning among treatment-seeking patients with chronic pain. *The Journal of Pain* **16** 844–851.

GOESLING, J., MOSER, S. E., ZAIDI, B., HASSETT, A. L., HILLIARD, P., HALLSTROM, B., CLAUW, D. J. and BRUMMETT, C. M. (2016). Trends and predictors of opioid use following total knee and total hip arthroplasty. *Pain* **157** 1259–1275.

HAH, J. M., SHARIFZADEH, Y., WANG, B. M., GILLESPIE, M. J., GOODMAN, S. B., MACKEY, S. C. and CARROLL, I. R. (2015). Factors associated with opioid use in a cohort of patients presenting for surgery. *Pain Research and Treatment* **2015**.

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

HILLIARD, P. E., WALJEE, J., MOSER, S., METZ, L., MATHIS, M., GOESLING, J., CRON, D., CLAUW, D. J., ENGLESBE, M. et al. (2018). Prevalence of preoperative opioid use and characteristics associated with opioid use among patients presenting for surgery. *JAMA Surg.* **153** 929–937.

JAIN, N., BROCK, J. L., PHILLIPS, F. M., WEAVER, T. and KHAN, S. N. (2018). Chronic preoperative opioid use is a risk factor for increased complications, resource use, and costs after cervical fusion. *Spine J.* **18** 1989–1998. https://doi.org/10.1016/j.spinee.2018.03.015

JANDA, A. M., AS-SANIE, S., RAJALA, B., TSODIKOV, A., MOSER, S. E., CLAUW, D. J. and BRUMMETT, C. M. (2015). Fibromyalgia survey criteria are associated with increased postoperative opioid consumption in women undergoing hysterectomy. *Anesthesiology*: *The Journal of the American Society of Anesthesiologists* **122** 1103–1111.

KARLIK, B. and OLGAC, A. V. (2011). Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems* **1** 111–122.

KLEESIEK, J., URBAN, G., HUBERT, A., SCHWARZ, D., MAIER-HEIN, K., BENDSZUS, M. and BILLER, A. (2016). Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *NeuroImage* **129** 460–469.

LECUN, Y., BENGIO, Y. and HINTON, G. (2015). Deep learning. *Nature* **521** 436–444.

LEE, D., ARMAGHANI, S., ARCHER, K. R., BIBLE, J., SHAU, D., KAY, H., ZHANG, C., MCGIRT, M. J. and DEVIN, C. (2014). Preoperative opioid use as a predictor of adverse postoperative self-reported outcomes in patients undergoing spine surgery. *JBJS* **96** e89(1)–e89(8).

LI, Y. and YUAN, Y. (2017). Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems* 597–607.

LIPTON, Z. C., KALE, D. C., ELKAN, C. and WETZEL, R. (2015). Learning to diagnose with LSTM recurrent neural networks. ArXiv preprint. Available at arXiv:1511.03677.

LIU, L., JIANG, H., HE, P., CHEN, W., LIU, X., GAO, J. and HAN, J. (2020). On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*.

LO-CIGANIC, W.-H., HUANG, J. L., ZHANG, H. H., WEISS, J. C., WU, Y., KWOH, C. K., DONOHUE, J. M., COCHRAN, G., GORDON, A. J. et al. (2019). Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. *JAMA Network Open* **2** e190968(1)–e190968(15).

MEREDITH, S. J., NADARAJAH, V., JAUREGUI, J. J., SMUDA, M. P., MEDINA, S. H., BENNETT, C. H., PACKER, J. D. and HENN III, R. F. (2019). Preoperative opioid use in knee surgery patients. *The Journal of Knee Surgery* **32** 630–636.

MORRIS, B. J., SCIASCIA, A. D., JACOBS, C. A. and EDWARDS, T. B. (2016). Preoperative opioid use associated with worse outcomes after anatomic shoulder arthroplasty. *Journal of Shoulder and Elbow Surgery* **25** 619–623.

PARK, B. U., MAMMEN, E., LEE, Y. K. and LEE, E. R. (2015). Varying coefficient regression models: A review and new developments. *Int. Stat. Rev.* **83** 36–64. MR3341079 https://doi.org/10.1111/insr.12029

PRENTICE, H. A., INACIO, M. C., SINGH, A., NAMBA, R. S. and PAXTON, E. W. (2019). Preoperative risk factors for opioid utilization after total hip arthroplasty. *JBJS* **101** 1670–1678.

ROECKEL, L.-A., COZ, G.-M. L., GAVÉRIAUX-RUFF, C. and SIMONIN, F. (2016). Opioid-induced hyperalgesia: Cellular and molecular mechanisms. *Neuroscience* **338** 160–182. https://doi.org/10.1016/j.neuroscience.2016.06.029

SAHA, T. D., KERRIDGE, B. T., GOLDSTEIN, R. B., CHOU, S. P., ZHANG, H., JUNG, J., PICKERING, R. P., RUAN, W. J., SMITH, S. M. et al. (2016). Nonmedical prescription opioid use and DSM-5 nonmedical prescription opioid use disorder in the United States. *J. Clin. Psychiatry* **77** 772–780.

SCHMIDT-HIEBER, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.* **48** 1875–1897. MR4134774 https://doi.org/10.1214/19-AOS1875

SCHOENFELD, A. J., BELMONT JR, P. J., BLUCHER, J. A., JIANG, W., CHAUDHARY, M. A., KOEHLMOOS, T., KANG, J. D. and HAIDER, A. H. (2018). Sustained preoperative opioid use is a predictor of continued use following spine surgery. *JBJS* **100** 914–921.

SEKHRI, S., ARORA, N. S., COTTRELL, H., BAERG, T., DUNCAN, A., HU, H. M., ENGLESBE, M. J., BRUMMETT, C. and WALJEE, J. F. (2018). Probability of opioid prescription refilling after surgery: Does initial prescription dose matter? *Ann. Surg.* **268** 271–276.

SHRESTHA, A. and MAHMOOD, A. (2019). Review of deep learning algorithms and architectures. *IEEE Access* **7** 53040–53065.

SMITH, S. R., BIDO, J., COLLINS, J. E., YANG, H., KATZ, J. N. and LOSINA, E. (2017). Impact of preoperative opioid use on total knee arthroplasty outcomes. *J. Bone Jt. Surg.* **99** 803–808.

SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I. and SALAKHUTDINOV, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15** 1929–1958. MR3231592

SULLIVAN, M. D., EDLUND, M. J., STEFFICK, D. and UNÜTZER, J. (2005). Regular use of prescribed opioids: Association with common psychiatric disorders. *Pain* **119** 95–103. https://doi.org/10.1016/j.pain.2005.09.020

SULLIVAN, M. D., EDLUND, M. J., ZHANG, L., UNÜTZER, J. and WELLS, K. B. (2006). Association between mental health disorders, problem drug use, and regular prescription opioid use. *Arch. Intern. Med.* **166** 2087–2093.

SUN, Y., KANG, J., BRUMMETT, C. and LI, Y. (2023). Supplement to "Individualized risk assessment of preoperative opioid use by interpretable neural network regression." https://doi.org/10.1214/22-AOAS1634SUPPA, https://doi.org/10.1214/22-AOAS1634SUPPB

TAN, G., JENSEN, M. P., THORNBY, J. I. and SHANTI, B. F. (2004). Validation of the brief pain inventory for chronic nonmalignant pain. *The Journal of Pain* **5** 133–137.

TRAN, P. V. (2016). A fully convolutional neural network for cardiac segmentation in short-axis MRI. ArXiv preprint. Available at arXiv:1604.00494.

TRAN, M.-N., NGUYEN, N., NOTT, D. and KOHN, R. (2020). Bayesian deep net GLM and GLMM. *J. Comput. Graph. Statist.* **29** 97–113. MR4085867 https://doi.org/10.1080/10618600.2019.1637747

WALJEE, J., CRON, D., STEIGER, R., ZHONG, L., ENGLESBE, M. and BRUMMETT, C. (2017). The effect of preoperative opioid exposure on healthcare utilization and expenditures following elective abdominal surgery. *Ann. Surg.* **265** 715–721.

WESTERMANN, R. W., HU, J., HAGEN, M. S., WILLEY, M., LYNCH, T. S. and ROSNECK, J. (2018). Epidemiology and detrimental impact of opioid use in patients undergoing arthroscopic treatment of femoroacetabular impingement syndrome. *Arthroscopy: The Journal of Arthroscopic & Related Surgery* **34** 2832–2836.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 454–475
https://doi.org/10.1214/22-AOAS1635

STATISTICAL MATCHING AND SUBCLASSIFICATION WITH A CONTINUOUS DOSE: CHARACTERIZATION, ALGORITHM, AND APPLICATION TO A HEALTH OUTCOMES STUDY

BY BO ZHANG[1,a], EMILY J. MACKAY[2,b] AND MIKE BAIOCCHI[3,c]

[1]*Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center,* [a]*bzhang3@fredhutch.org*
[2]*Department of Anesthesiology and Critical Care, Perelman School of Medicine, University of Pennsylvania,* [b]*Emily.Mackay@pennmedicine.upenn.edu*
[3]*Department of Epidemiology and Population Health, Stanford University,* [c]*baiocchi@stanford.edu*

Subclassification and matching are often used in empirical studies to adjust for observed covariates; however, they are largely restricted to relatively simple study designs with a binary treatment and less developed for designs with a continuous exposure. Matching with exposure doses is particularly useful in instrumental variable designs and in understanding the dose-response relationships. In this article we propose two criteria for optimal subclassification based on subclass homogeneity, in the context of having a continuous exposure dose, and propose an efficient polynomial-time algorithm that is guaranteed to find an optimal subclassification with respect to one criterion and serves as a 2-approximation algorithm for the other criterion. We discuss how to incorporate dose and use appropriate penalties to control the number of subclasses in the design. Via extensive simulations, we systematically compare our proposed design to optimal nonbipartite pair matching and demonstrate that combining our proposed subclassification scheme with regression adjustment helps to reduce model dependence for parametric causal inference with a continuous dose. We apply the new design and associated randomization-based inferential procedure to study the effect of transesophageal echocardiography (TEE) monitoring during coronary artery bypass graft (CABG) surgery on patients' postsurgery clinical outcomes, using Medicare and Medicaid claims data, and find evidence that TEE monitoring lowers patients' all-cause 30-day mortality rate.

REFERENCES

BAIOCCHI, M., SMALL, D. S., LORCH, S. and ROSENBAUM, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Amer. Statist. Assoc.* **105** 1285–1296. MR2796550 https://doi.org/10.1198/jasa.2010.ap09490

BAIOCCHI, M., SMALL, D. S., YANG, L., POLSKY, D. and GROENEVELD, P. W. (2012). Near/far matching: A study design approach to instrumental variables. *Health Serv. Outcomes Res. Methodol.* **12** 237–253.

COCHRAN, W. G. and RUBIN, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya, Ser. A* 417–446.

COLE, S. R. and STUART, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Am. J. Epidemiol.* **172** 107–115.

DING, P., FELLER, A. and MIRATRIX, L. (2016). Randomization inference for treatment effect variation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 655–671. MR3506797 https://doi.org/10.1111/rssb.12124

HAHN, R. T., ABRAHAM, T., ADAMS, M. S., BRUCE, C. J., GLAS, K. E., LANG, R. M., REEVES, S. T., SHANEWISE, J. S., SIU, S. C. et al. (2013). Guidelines for performing a comprehensive transesophageal echocardiographic examination: Recommendations from the American society of echocardiography and the society of cardiovascular anesthesiologists. *J. Am. Soc. Echocardiogr.* **26** 921–964.

HANSEN, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist. Assoc.* **99** 609–618. MR2086387 https://doi.org/10.1198/016214504000000647

HANSEN, B. B. (2007). Optmatch: Flexible, optimal matching for observational studies. *R News* **7** 18–24.

Key words and phrases. Approximation algorithm, continuous dose, dose-response relationship, edge cover, full match, subclassification.

HANSEN, B. B. and KLOPFER, S. O. (2006). Optimal full matching and related designs via network flows. *J. Comput. Graph. Statist.* **15** 609–627. MR2280151 https://doi.org/10.1198/106186006X137047

HENG, S., KANG, H., SMALL, D. S. and FOGARTY, C. B. (2021). Increasing power for observational studies of aberrant response: An adaptive approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 482–504. MR4294541 https://doi.org/10.1111/rssb.12424

HENG, S., ZHANG, B., HAN, X., LORCH, S. A. and SMALL, D. S. (2019). Instrumental variables: To strengthen or not to strengthen? ArXiv preprint. Available at arXiv:1911.09171.

HILLIS, L. D., SMITH, P. K., ANDERSON, J. L., BITTL, J. A., BRIDGES, C. R., BYRNE, J. G., CIGARROA, J. E., DISESA, V. J., HIRATZKA, L. F. et al. (2011). 2011 ACCF/AHA guideline for coronary artery bypass graft surgery: A report of the American college of cardiology foundation/American heart association task force on practice guidelines developed in collaboration with the American association for thoracic surgery, society of cardiovascular anesthesiologists, and society of thoracic surgeons. *J. Am. Coll. Cardiol.* **58** e123–e210.

HIRANO, K. and IMBENS, G. W. (2004). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. *Wiley Ser. Probab. Stat.* 73–84. Wiley, Chichester. MR2134803 https://doi.org/10.1002/0470090456.ch7

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* **15** 199–236.

IMBENS, G. W. and ROSENBAUM, P. R. (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *J. Roy. Statist. Soc. Ser. A* **168** 109–126. MR2113230 https://doi.org/10.1111/j.1467-985X.2004.00339.x

LOPEZ, M. J. and GUTMAN, R. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. *Statist. Sci.* **32** 432–454. MR3696004 https://doi.org/10.1214/17-STS612

LU, B., GREEVY, R., XU, X. and BECK, C. (2011). Optimal nonbipartite matching and its statistical applications. *Amer. Statist.* **65** 21–30. MR2899649 https://doi.org/10.1198/tast.2011.08294

LU, B., ZANUTTO, E., HORNIK, R. and ROSENBAUM, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *J. Amer. Statist. Assoc.* **96** 1245–1253. MR1973668 https://doi.org/10.1198/016214501753381896

MACKAY, E. J., NEUMAN, M. D., FLEISHER, L. A., PATEL, P. A., GUTSCHE, J. T., AUGOUSTIDES, J. G., DESAI, N. D. and GROENEVELD, P. W. (2020a). Transesophageal echocardiography, mortality, and length of hospitalization after cardiac valve surgery. *J. Am. Soc. Echocardiogr.*

MACKAY, E. J., ZHANG, B., HENG, S. and YE, T. (2020b). Protocol for a retrospective, comparative effectiveness study of the association between transesophageal echocardiography (TEE) monitoring used in Coronary Artery Bypass Graft (CABG) surgery and clinical outcomes. *MedRxiv*.

MACKAY, E. J., ZHANG, B., HENG, S., YE, T., NEUMAN, M. D., AUGOUSTIDES, J. G., FEINMAN, J. W., DESAI, N. D. and GROENEVELD, P. W. (2021). Association between transesophageal echocardiography and clinical outcomes after coronary artery bypass graft surgery. *J. Am. Soc. Echocardiogr.* **34** 571–581.

NISHIMURA, R. A., OTTO, C. M., BONOW, R. O., CARABELLO, B. A., ERWIN, J. P., FLEISHER, L. A., JNEID, H., MACK, M. J., MCLEOD, C. J. et al. (2017). 2017 AHA/ACC focused update of the 2014 AHA/ACC guideline for the management of patients with valvular heart disease: A report of the American college of cardiology/American heart association task force on clinical practice guidelines. *J. Am. Coll. Cardiol.* **70** 252–289.

THE SOCIETY OF THORACIC SURGEONS (2016). The STS Adult Cardiac Surgery Database (ACSD).

ROSENBAUM, P. R. (1989). Optimal matching for observational studies. *J. Amer. Statist. Assoc.* **84** 1024–1032.

ROSENBAUM, P. R. (1991). A characterization of optimal designs for observational studies. *J. Roy. Statist. Soc. Ser. B* **53** 597–610. MR1125717

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. (2004). Design sensitivity in observational studies. *Biometrika* **91** 153–164. MR2050466 https://doi.org/10.1093/biomet/91.1.153

ROSENBAUM, P. R. (2010a). *Design of Observational Studies*. *Springer Series in Statistics*. Springer, New York. MR2561612 https://doi.org/10.1007/978-1-4419-1213-8

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

RUBIN, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* 159–183.

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–328.

RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *J. Amer. Statist. Assoc.* **81** 961–962.

RUBIN, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat. Med.* **26** 20–36. MR2312697 https://doi.org/10.1002/sim.2739

SCHRIJVER, A. (2003). *Combinatorial Optimization. Polyhedra and Efficiency. Vol. A. Algorithms and Combinatorics* **24**. Springer, Berlin. MR1956924

SMALL, D. S. and ROSENBAUM, P. R. (2008). War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *J. Amer. Statist. Assoc.* **103** 924–933. MR2528819 https://doi.org/10.1198/016214507000001247

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313

VAZIRANI, V. V. (2001). *Approximation Algorithms*. Springer, Berlin.

WILLIAMSON, D. P. and SHMOYS, D. B. (2011). *The Design of Approximation Algorithms*. Cambridge Univ. Press, Cambridge. MR2798112 https://doi.org/10.1017/CBO9780511921735

WU, X., MEALLI, F., KIOUMOURTZOGLOU, M.-A., DOMINICI, F. and BRAUN, D. (2018). Matching on generalized propensity scores with continuous exposures. ArXiv preprint. Available at arXiv:1812.06575.

YANG, S., IMBENS, G. W., CUI, Z., FARIES, D. E. and KADZIOLA, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics* **72** 1055–1065. MR3591590 https://doi.org/10.1111/biom.12505

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

ZHANG, B., HENG, S., MACKAY, E. J. and YE, T. (2021a). Bridging preference-based instrumental variable studies and cluster-randomized encouragement experiments: Study design, noncompliance, and average cluster effect ratio. *Biometrics* (in press).

ZHANG, B., MACKAY, E. J. and BAIOCCHI, M. (2023). Supplement to “Statistical matching and subclassification with a continuous dose: Characterization, algorithm, and application to a health outcomes study.” https://doi.org/10.1214/22-AOAS1635SUPPA, https://doi.org/10.1214/22-AOAS1635SUPPB

ZHANG, B., SMALL, D. S., LASATER, K. B., MCHUGH, M., SILBER, J. H. and ROSENBAUM, P. R. (2021b). Matching one sample according to two criteria in observational studies. *J. Amer. Statist. Assoc.* **0** 1–12.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 476–497
https://doi.org/10.1214/22-AOAS1638

ESTIMATING PROMOTION EFFECTS IN EMAIL MARKETING USING A LARGE-SCALE CROSS-CLASSIFIED BAYESIAN JOINT MODEL FOR NESTED IMBALANCED DATA

BY SABYASACHI MUKHOPADHYAY[1,a], WREETABRATA KAR[2,b] AND GOURAB MUKHERJEE[3,c]

[1]*Department of Quantitative Methods and Information Systems, Indian Institute of Management,* [a]*sabyasachi.mukhopadhyay@iimu.ac.in*

[2]*Krannert School of Management, Purdue University,* [b]*wkar@purdue.edu*

[3]*Department of Data Sciences and Operations, University of Southern California,* [c]*gourab@usc.edu*

We consider a large-scale, *cross-classified nested* (CRON) joint model for modeling customer responses to opening, clicking, and purchasing from promotion emails. Our logistic regression-based joint model contains crossing of promotions and customer effects and allows estimation of the heterogeneous effects of different promotion emails, after adjusting for customer preferences, attributes, and historical behaviors. Using data from an email marketing campaign of an apparel retailer, we exhibit the varying effects of promotions not only based on the contents of the email but also across the three different stages, viz. open, click, and purchase of the conversion funnel. We conduct Bayesian estimation of the parameters in the joint model by using a block Metropolis–Hastings algorithm that not only incorporates nested subsampling to tackle the severe imbalance between conversions and no conversions but also uses additive transformation-based modifications of random walk Metropolis to scale estimation for large numbers of customers. We extend our approach to a *segmented cross-classified nested* (SCRON) joint model that encompasses the possibility of varying promotion effects across different customer segments. The resultant high-dimensional model is estimated using spike-and-slab priors on the promotion and customer segment interactions. Our nested joint model accounts for the correlations in customer preferences across the conversion funnel. Based on the promotion estimates from the model, we demonstrate how marketers can use different priced, nonpriced, and combination of price and nonprice promotions to increase brand awareness or increase purchases. Comparing estimates from CRON and SCRON models, we display the benefits of targeted marketing by using email promotion lists which are separately optimized for the different customer segments.

REFERENCES

ALDIGHIERI, R. (2019). Marketer email tracker 2019. https://dma.org.uk/uploads/misc/marketers-email-tracker-2019.pdf.

ALQUIER, P., FRIEL, N., EVERITT, R. and BOLAND, A. (2016). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Stat. Comput.* **26** 29–47. MR3439357 https://doi.org/10.1007/s11222-014-9521-x

ANDERSON, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59** 19–35. MR0345332 https://doi.org/10.1093/biomet/59.1.19

BAI, R., ROCKOVA, V. and GEORGE, E. I. (2021). Spike-and-slab meets LASSO: A review of the spike-and-slab LASSO. In *Handbook of Bayesian Variable Selection* 81–108.

BISSIRI, P. G., HOLMES, C. C. and WALKER, S. G. (2016). A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 1103–1130. MR3557191 https://doi.org/10.1111/rssb.12158

BULT, J. R. and WANSBEEK, T. (1995). Optimal selection for direct mail. *Mark. Sci.* **14** 378–394.

Key words and phrases. Logistic regression, joint modeling, imbalanced data, spike-and-slab priors, Bayesian hierarchical model, promotion email, high-dimensional regression.

Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874 https://doi.org/10.1214/15-AOS1334

Chawla, N. V., Japkowicz, N. and Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD Explor. Newsl.* **6** 1–6.

Dey, K. K. and Bhattacharya, S. (2019). A brief review of optimal scaling of the main MCMC approaches and optimal scaling of additive TMCMC under non-regular cases. *Braz. J. Probab. Stat.* **33** 222–266. MR3919022 https://doi.org/10.1214/17-bjps386

Dey, K. Kr. and Bhattacharya, S. (2016). On geometric ergodicity of additive and multiplicative transformation-based Markov Chain Monte Carlo in high dimensions. *Braz. J. Probab. Stat.* **30** 570–613. MR3582391 https://doi.org/10.1214/15-BJPS295

Dutta, S. and Bhattacharya, S. (2014). Markov chain Monte Carlo based on deterministic transformations. *Stat. Methodol.* **16** 100–116. MR3110892 https://doi.org/10.1016/j.stamet.2013.08.006

Fader, P. S., Hardie, B. G. and Lee, K. L. (2005). Counting your customers the easy way: An alternative to the Pareto/NBD model. *Mark. Sci.* **24** 275–284.

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

Forte, D. (2019). Consumers want more personalized emails, but retailers aren't delivering. Available at https://multichannelmerchant.com/marketing/consumers-want-personalized-emails-retailers-arent-delivering/.

Gao, K. (2017). *Scalable Estimation and Inference for Massive Linear Mixed Models with Crossed Random Effects*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Stanford University. MR4257222

Gao, K. and Owen, A. (2017). Efficient moment calculations for variance components in large unbalanced crossed random effects models. *Electron. J. Stat.* **11** 1235–1296. MR3635913 https://doi.org/10.1214/17-EJS1236

Geisser, S. (1993). *Predictive Inference: An Introduction. Monographs on Statistics and Applied Probability* **55**. CRC Press, New York. MR1252174 https://doi.org/10.1007/978-1-4899-4467-2

Gopalakrishnan, A. and Park, Y.-H. (2021). The impact of coupons on the visit-to-purchase funnel: Theory and empirical evidence. *Mark. Sci.*.

Huang, Z. and Gelman, A. (2005). Sampling for Bayesian computation with large datasets. Available at SSRN 1010107.

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

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning. Springer Texts in Statistics* **103**. Springer, New York. With applications in R. MR3100153 https://doi.org/10.1007/978-1-4614-7138-7

Johndrow, J. E., Smith, A., Pillai, N. and Dunson, D. B. (2019). MCMC for imbalanced categorical data. *J. Amer. Statist. Assoc.* **114** 1394–1403. MR4011787 https://doi.org/10.1080/01621459.2018.1505626

Kamakura, W. A. and Kang, W. (2007). Chain-wide and store-level analysis for cross-category management. *J. Retail.* **83** 159–170.

Kapner, S. (2017). Retailers' emails are misfires for many holiday shoppers.

Kleijn, B. J. K. and Van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electron. J. Stat.* **6** 354–381.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer, New York. MR3099297 https://doi.org/10.1007/978-1-4614-6849-3

Lee (2019). Digital vs traditional media—where should you invest your marketing dollars? Available at https://hookdpromotions.com/digital-vs-traditional-media-where-should-you-invest-your-marketing-dollars/.

Leone, C. (2020). How much should you budget for marketing in 2020. Available at https://www.webstrategiesinc.com/blog/how-much-budget-for-misc-marketing-in-2014.

Malsiner-Walli, G. and Wagner, H. (2011). Comparing spike and slab priors for Bayesian variable selection. *Aust. J. Stat.*

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R. and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat. Med.* **32** 3388–3414. MR3074364 https://doi.org/10.1002/sim.5753

McCulloch, C. (2008). Joint modelling of mixed outcome types using latent variables. *Stat. Methods Med. Res.* **17** 53–73. MR2420190 https://doi.org/10.1177/0962280207081240

Moore, D. J. and Lee, S. P. (2012). How advertising influences consumption impulses. *J. Advert.* **41** 107–120.

Mukhopadhyay, S., Kar, W. and Mukherjee, G. (2023). Supplement to "Estimating promotion effects in email marketing using a large-scale cross-classified Bayesian joint model for nested imbalanced data." https://doi.org/10.1214/22-AOAS1638SUPP

Osinga, E. C., Leeflang, P. S. and Wieringa, J. E. (2010). Early marketing matters: A time-varying parameter approach to persistence modeling. *J. Mark. Res.* **47** 173–185.

OWEN, A. B. (2007). Infinitely imbalanced logistic regression. *J. Mach. Learn. Res.* **8** 761–773. MR2320678

PAPASPILIOPOULOS, O., ROBERTS, G. O. and ZANELLA, G. (2020). Scalable inference for crossed random effects models. *Biometrika* **107** 25–40. MR4064138 https://doi.org/10.1093/biomet/asz058

PARK, C. H., PARK, Y.-H. and SCHWEIDEL, D. A. (2018). The effects of mobile promotions on customer purchase dynamics. *Int. J. Res. Mark.* **35** 453–470.

PERRIN, N. (2019). Email marketing 2019: Still a leading touchpoint for marketers and consumers alike. Available at: https://content-na1.emarketer.com/email-marketing-2019.

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

PRENTICE, R. L. and PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–411. MR0556730 https://doi.org/10.1093/biomet/66.3.403

RIZOPOULOS, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press, Boca Raton.

RIZOPOULOS, D. and GHOSH, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat. Med.* **30** 1366–1380. MR2828959 https://doi.org/10.1002/sim.4205

RIZOPOULOS, D. and LESAFFRE, E. (2014). Introduction to the special issue on joint modelling techniques. *Stat. Methods Med. Res.* **23** 3–10. MR3190684 https://doi.org/10.1177/0962280212445800

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

ROSSI, P. E., MCCULLOCH, R. E. and ALLENBY, G. M. (1996). The value of purchase history data in target marketing. *Mark. Sci.* **15** 321–340.

RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.* **11** 735–757. MR1944261 https://doi.org/10.1198/106186002321018768

SACHS, M., SEN, D., LU, J. and DUNSON, D. (2020). Posterior computation with the gibbs zig-zag sampler. Preprint. Available at arXiv:2004.04254.

SAHNI, N. S., WHEELER, S. C. and CHINTAGUNTA, P. (2018). Personalization in email marketing: The role of noninformative advertising content. *Mark. Sci.* **37** 236–258.

SAHNI, N. S., ZOU, D. and CHINTAGUNTA, P. K. (2017). Do targeted discount offers serve as advertising? Evidence from 70 field experiments. *Manage. Sci.* **63** 2688–2705.

SCOTT, S. L., BLOCKER, A. W., BONASSI, F. V., CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *Int. J. Manag. Sci. Eng. Manag.* **11** 78–88.

SEETHARAMAN, P. B. and CHINTAGUNTA, P. K. (2003). The proportional hazard model for purchase timing: A comparison of alternative specifications. *J. Bus. Econom. Statist.* **21** 368–382. MR1982528 https://doi.org/10.1198/073500103288619025

SEN, D., SACHS, M., LU, J. and DUNSON, D. B. (2020). Efficient posterior sampling for high-dimensional imbalanced logistic regression. *Biometrika* **107** 1005–1012. MR4186502 https://doi.org/10.1093/biomet/asaa035

STATISTA (2019). Dossier on e-mail marketing in the U.S. and worldwide.

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

WU, J., LI, K. J. and LIU, J. S. (2018). Bayesian inference for assessing effects of email marketing campaigns. *J. Bus. Econom. Statist.* **36** 253–266. MR3790212 https://doi.org/10.1080/07350015.2016.1141096

ZHANG, X., KUMAR, V. and COSGUNER, K. (2017). Dynamically managing a profitable email marketing program. *J. Mark. Res.* **54** 851–866.

ZHANG, X., ZHOU, Y., MA, Y., CHEN, B.-C., ZHANG, L. and AGARWAL, D. (2016). Glmix: Generalized linear mixed models for large-scale response prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 363–372.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 498–517
https://doi.org/10.1214/22-AOAS1639

MODELING PANELS OF EXTREMES

BY DEBBIE J. DUPUIS[1,a], SEBASTIAN ENGELKE[2,b] AND LUCA TRAPIN[3,c]

[1]*Department of Decision Sciences, HEC Montréal,* [a]*debbie.dupuis@hec.ca*
[2]*Research Center for Statistics, University of Geneva,* [b]*sebastian.engelke@unige.ch*
[3]*Department of Statistics, University of Bologna,* [c]*luca.trapin@unibo.it*

Extreme value applications commonly employ regression techniques to capture cross-sectional heterogeneity or time variation in the data. Estimation of the parameters of an extreme value regression model is notoriously challenging due to the small number of observations that are usually available in applications. When repeated extreme measurements are collected on the same individuals, that is, a panel of extremes is available, pooling the observations in groups can improve the statistical inference. We study three data sets related to risk assessment in finance, climate science, and hydrology. In all three cases the problem can be formulated as an extreme value panel regression model with a latent group structure and group-specific parameters. We propose a new algorithm that jointly assigns the individuals to the latent groups and estimates the parameters of the regression model inside each group. Our method efficiently recovers the underlying group structure without prior information, and for the three data sets it provides improved return level estimates and helps answer important domain-specific questions.

REFERENCES

ALILA, Y. (1999). A hierarchical approach for the regionalization of precipitation annual maxima in Canada. *J. Geophys. Res.* **104** 31,645–31,655.

ASADI, P., DAVISON, A. C. and ENGELKE, S. (2015). Extremes on river networks. *Ann. Appl. Stat.* **9** 2023–2050. MR3456363 https://doi.org/10.1214/15-AOAS863

ASADI, P., ENGELKE, S. and DAVISON, A. C. (2018). Optimal regionalization of extreme value distributions for flood estimation. *J. Hydrol.* **556** 182–193.

BALI, T. G. (2003). An extreme value approach to estimating volatility and value at risk. *J. Bus.* **76** 83–108.

BARNDORFF-NIELSEN, O. E., KINNEBROCK, S. and SHEPHARD, N. (2010). Measuring downside risk—realized semivariance. In *Volatility and Time Series Econometrics*: *Essays in Honor of Robert F. Engle*. *Adv. Texts Econometrics* 117–136. Oxford Univ. Press, Oxford. MR3237181 https://doi.org/10.1093/acprof:oso/9780199549498.003.0007

BEE, M., DUPUIS, D. J. and TRAPIN, L. (2019). Realized peaks over threshold: A time-varying extreme value approach with high-frequency-based measures. *J. Financ. Econom.* **17** 254–283.

BOCCHIOLA, D., BIANCHI JANETTI, E., GORNI, E., MARTY, C. and SOVILLA, B. (2008). Regional evaluation of three-day snow depth for avalanche hazard mapping in Switzerland. *Nat. Hazards Earth Syst. Sci.* **8** 685–705.

BÜCHER, A. and SEGERS, J. (2017). On the maximum likelihood estimator for the generalized extreme-value distribution. *Extremes* **20** 839–872. MR3737387 https://doi.org/10.1007/s10687-017-0292-6

BÜCHER, A. and SEGERS, J. (2018). Maximum likelihood estimation for the Fréchet distribution based on block maxima extracted from a time series. *Bernoulli* **24** 1427–1462. MR3706798 https://doi.org/10.3150/16-BEJ903

BUISHAND, T. A. (1991). Extreme rainfall estimation by combining data from several sites. *Hydrol. Sci. J.* **36** 345–365.

BURN, D. H. (1990). An appraisal of the "region of influence" approach to flood frequency analysis. *Hydrol. Sci. J.* **35** 149–165.

CAMERON, C. A. and TRIVEDI, P. K. (2015). Count panel data. In *The Oxford Handbook of Panel Data* (B. H. Baltagi, ed.) 8. Oxford Univ. Press, Oxford.

CANNON, A. J. (2010). A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrol. Process.* **24** 673–685.

Key words and phrases. Clustering, extreme value regression, panel data, risk assessment.

CARREAU, J., NAVEAU, P. and NEPPEL, L. (2017). Partitioning into hazard subregions for regional peaks-over-threshold modeling of heavy precipitation. *Water Resour. Res.* **53** 4407–4426. https://doi.org/10.1002/2017WR020758

CHANDLER, R. E. and BATE, S. (2007). Inference for clustered data using the independence loglikelihood. *Biometrika* **94** 167–183. MR2367830 https://doi.org/10.1093/biomet/asm015

CHAVEZ-DEMOULIN, V. and DAVISON, A. C. (2005). Generalized additive modelling of sample extremes. *J. R. Stat. Soc. Ser. C.* **54** 207–222. MR2134607 https://doi.org/10.1111/j.1467-9876.2005.00479.x

CHENG, E. D. H. (1998). Macroscopic extreme wind regionalization. *J. Wind Eng. Ind. Aerodyn.* **77-78** 13–21.

COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. *Springer Series in Statistics*. Springer, London. MR1932132 https://doi.org/10.1007/978-1-4471-3675-0

COMMITTEE ON THE GLOBAL FINANCIAL SYSTEM (2001). Stress testing by large financial institutions: Current practice and aggregation issues Technical Report Bank for International Settlements.

DAVISON, A. C., PADOAN, S. A. and RIBATET, M. (2012). Statistical modeling of spatial extremes. *Statist. Sci.* **27** 161–186. MR2963980 https://doi.org/10.1214/11-STS376

DE CARVALHO, M. and DAVISON, A. C. (2014). Spectral density ratio models for multivariate extremes. *J. Amer. Statist. Assoc.* **109** 764–776. MR3223748 https://doi.org/10.1080/01621459.2013.872651

DE CARVALHO, M., PEREIRA, S., PEREIRA, P. and DE ZEA BERMUDEZ, P. (2022). An extreme value Bayesian Lasso for the conditional left and right tails. *J. Agric. Biol. Environ. Stat.* **27** 222–239. MR4416781 https://doi.org/10.1007/s13253-021-00469-9

DOMBRY, C. (2015). Existence and consistency of the maximum likelihood estimators for the extreme value index within the block maxima framework. *Bernoulli* **21** 420–436. MR3322325 https://doi.org/10.3150/13-BEJ573

DOMBRY, C. and FERREIRA, A. (2019). Maximum likelihood estimators based on the block maxima method. *Bernoulli* **25** 1690–1723. MR3961227 https://doi.org/10.3150/18-BEJ1032

DUPUIS, D. J, ENGELKE, S. and TRAPIN, L. (2023). Supplement to “Modeling panels of extremes.” https://doi.org/10.1214/22-AOAS1639SUPP

EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (1997). *Modelling Extremal Events*: *For Insurance and Finance*. *Applications of Mathematics* (*New York*) **33**. Springer, Berlin. MR1458613 https://doi.org/10.1007/978-3-642-33483-2

ENGELKE, S. and HITZ, A. S. (2020). Graphical models for extremes (with discussions). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 871–932. MR4136498

GARCÍA, G., DRECCER, M., MIRALLES, D. and SERRAGO, R. (2015). High night temperatures during grain number determination reduce wheat and barley grain yield: A field study. *Glob. Change Biol.* **21** 4153–4164.

GNECCO, N., TEREFE, E. M. and ENGELKE, S. (2022). Extremal Random Forests. https://doi.org/10.48550/ARXIV.2201.12865

GNECCO, N., MEINSHAUSEN, N., PETERS, J. and ENGELKE, S. (2021). Causal discovery in heavy-tailed models. *Ann. Statist.* **49** 1755–1778. MR4298880 https://doi.org/10.1214/20-aos2021

GREENE, W. (2009). Discrete choice modeling. In *Palgrave Handbook of Econometrics* 473–556. Springer, Berlin.

GU, J. and VOLGUSHEV, S. (2019). Panel data quantile regression with grouped fixed effects. *J. Econometrics* **213** 68–91. MR4013216 https://doi.org/10.1016/j.jeconom.2019.04.006

HAMBUCKERS, J. and KNEIB, T. (2021). Smooth transition regression models for non-stationary extremes. *J. Financ. Econom.* **nbab005**.

HANSEN, J., RUEDY, R., SATO, M. and LO, K. (2010). Global surface temperature change. *Rev. Geophys.* **48**. RG4004.

HSIAO, C. (2007). Panel data analysis—advantages and challenges. *TEST* **16** 1–22. MR2368444 https://doi.org/10.1007/s11749-007-0046-x

HSIAO, C. (2014). *Analysis of Panel Data*, 3rd ed. *Econometric Society Monographs* **54**. Cambridge Univ. Press, New York. MR3560738 https://doi.org/10.1017/CBO9781139839327

IPCC (2008). Climate change 2007. Synthesis report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change Core Writing Team, eds. R. K. Pachauri and A. Reisinger.

JURADO, K., LUDVIGSON, S. C. and NG, S. (2015). Measuring uncertainty. *Am. Econ. Rev.* **105** 1177–1216.

KATZ, R. W. (2013). Statistical methods for nonstationary extremes. In *Extremes in a Changing Climate*. *Water Sci. Technol. Libr.* **65** 15–37. Springer, Dordrecht. MR3051152 https://doi.org/10.1007/978-94-007-4479-0_2

KATZ, R. W., PARLANGE, M. B. and NAVEAU, P. (2002). Statistics of extremes in hydrology. *Adv. Water Resour.* **25** 1287–1304.

KUKAL, M. S. and IRMAK, S. (2018). Climate-driven crop yield and yield variability and climate change impacts on the U.S. Great Plains agricultural production. *Sci. Rep.* **8** 3450. https://doi.org/10.1038/s41598-018-21848-2

LEADBETTER, M. R., LINDGREN, G. and ROOTZÉN, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. *Springer Series in Statistics*. Springer, New York-Berlin. MR0691492

MASSACCI, D. (2017). Tail risk dynamics in stock returns: Links to the macroeconomy and global markets connectedness. *Manage*. *Sci*. **63** 3072–3089.

MCLACHLAN, G. J. and KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR2392878 https://doi.org/10.1002/9780470191613

MCNEIL, A. J., FREY, R. and EMBRECHTS, P. (2015). *Quantitative Risk Management*: *Concepts*, *Techniques and Tools*, Revised ed. *Princeton Series in Finance*. Princeton Univ. Press, Princeton, NJ. MR3445371

MERZ, R. and BLÖSCHL, G. (2005). Flood frequency regionalisation—spatial proximity vs. catchment attributes. *J*. *Hydrol*. **302** 283–306.

MHALLA, L., CHAVEZ-DEMOULIN, V. and DUPUIS, D. J. (2020). Causal mechanism of extreme river discharges in the upper Danube basin network. *J*. *R*. *Stat*. *Soc*. *Ser*. *C*. *Appl*. *Stat*. **69** 741–764. MR4133145

MHALLA, L., HAMBUCKERS, J. and LAMBERT, M. (2020). Extremal connectedness and systemic risk of hedge funds. Working Paper (Available at SSRN).

OH, D. H. and PATTON, A. J. (2020). Dynamic Factor Copula Models with Estimated Cluster Assignments. Working Paper.

OVEREEM, A., BUISHAND, A. and HOLLEMAN, I. (2008). Rainfall depth-duration-frequency curves and their uncertainties. *J*. *Hydrol*. **348** 124–134.

PAKEL, C., SHEPHARD, N. and SHEPPARD, K. (2011). Nuisance parameters, composite likelihoods and a panel of GARCH models. *Statist*. *Sinica* **21** 307–329. MR2796864

PAPALEXIOU, S. M., AGHAKOUCHAK, A., TRENBERTH, K. E. and FOUFOULA-GEORGIOU, E. (2018). Global, regional, and megacity trends in the highest temperature of the year: Diagnostics and evidence for accelerating trends. *Earths Future* **6** 71–79. https://doi.org/10.1002/2017EF000709

PETERSEN, L. K. (2019). Impact of climate change on twenty-first century crop yields in the U.S. *Climate* **7**.

PICKANDS, J. III (1975). Statistical inference using extreme order statistics. *Ann*. *Statist*. **3** 119–131. MR0423667

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J*. *Amer*. *Statist*. *Assoc*. **66** 846–850.

REICH, B. J. and SHABY, B. A. (2019). A spatial Markov model for climate extremes. *J*. *Comput*. *Graph*. *Statist*. **28** 117–126. MR3939376 https://doi.org/10.1080/10618600.2018.1482764

RIBATET, M. (2019). SpatialExtremes: Modelling Spatial Extremes. R package version 2.0-7.2.

ROHRBECK, C. and TAWN, J. A. (2021). Bayesian spatial clustering of extremal behavior for hydrological variables. *J*. *Comput*. *Graph*. *Statist*. **30** 91–105. MR4235967 https://doi.org/10.1080/10618600.2020.1777139

RÖTTGER, F., ENGELKE, S. and ZWIERNIK, P. (2021). Total positivity in multivariate extremes. Available from https://arxiv.org/abs/2112.14727.

SADOK, W. and KRISHNA JAGADISH, S. V. (2020). The hidden costs of nighttime warming on yields. *Trends Plant Sci*. **25** 644–651.

SEGAL, G., SHALIASTOVICH, I. and YARON, A. (2015). Good and bad uncertainty: Macroeconomic and financial market implications. *J*. *Financ*. *Econ*. **117** 369–397.

SMITH, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72** 67–90. MR0790201 https://doi.org/10.1093/biomet/72.1.67

SU, L., SHI, Z. and PHILLIPS, P. C. B. (2016). Identifying latent structures in panel data. *Econometrica* **84** 2215–2264. MR3580267 https://doi.org/10.3982/ECTA12560

VELTHOEN, J., DOMBRY, C., CAI, J.-J. and ENGELKE, S. (2021). Gradient boosting for extreme quantile regression. Available at arXiv:2103.00808.

VIGNOTTO, E., ENGELKE, S. and ZSCHEISCHLER, J. (2021). Clustering bivariate dependencies of compound precipitation and wind extremes over Great Britain and Ireland. *Weather Clim*. *Extrem*. **32**.

WANG, J., HAN, Y., STEIN, M. L., KOTAMARTHI, V. R. and HUANG, W. K. (2016). Evaluation of dynamically downscaled extreme temperature using a spatially-aggregated generalized extreme value (GEV) model. *Clim*. *Dyn*. **47** 2833–2848.

WANG, W. and SU, L. (2021). Identifying latent group structures in nonlinear panels. *J*. *Econometrics* **220** 272–295. MR4201491 https://doi.org/10.1016/j.jeconom.2020.04.003

ZHAO, Z., ZHANG, Z. and CHEN, R. (2018). Modeling maxima with autoregressive conditional Fréchet model. *J*. *Econometrics* **207** 325–351. MR3868186 https://doi.org/10.1016/j.jeconom.2018.07.004

ZWIERS, F. W. and KHARIN, V. V. (1998). Changes in the extremes of the climate simulated by CCC GCM2 under CO2 doubling. *J*. *Climate* **11** 2200–2222.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 518–538
https://doi.org/10.1214/22-AOAS1640

A VARIATIONAL BAYESIAN APPROACH TO IDENTIFYING WHOLE-BRAIN DIRECTED NETWORKS WITH FMRI DATA

BY YAOTIAN WANG[1,a], GUOFEN YAN[2,d], XIAOFENG WANG[3,e], SHUORAN LI[1,b], LINGYI PENG[4,f], DANA L. TUDORASCU[4,g] AND TINGTING ZHANG[1,c]

[1]*Department of Statistics, University of Pittsburgh,* [a]*yaw87@pitt.edu,* [b]*shl198@pitt.edu,* [c]*tiz67@pitt.edu*
[2]*Department of Public Health Sciences, University of Virginia,* [d]*guofen.yan@virginia.edu*
[3]*Department of Quantitative Health Sciences, Cleveland Clinic,* [e]*wangx6@ccf.org*
[4]*Department of Biostatistics, University of Pittsburgh,* [f]*lingyi.peng@pitt.edu,* [g]*dlt30@pitt.edu*

The brain is a high-dimensional directed network system, as it consists of many regions as network nodes that exert influence on each other. The directed influence exerted by one region on another is referred to as directed connectivity. We aim to reveal whole-brain directed networks based on resting-state functional magnetic resonance imaging (fMRI) data of many subjects. However, it is both statistically and computationally challenging to produce scientifically meaningful estimates of whole-brain directed networks. To address the statistical modeling challenge, we assume modular brain networks which reflect functional specialization and functional integration of the brain. We address the computational challenge by developing a variational Bayesian method to estimate the new model. We apply our method to resting-state fMRI data of many subjects and identify modules and directed connections in whole-brain directed networks. The identified modules are accordant with functional brain systems specialized for different functions. We also detect directed connections between functionally specialized modules, which is not attainable by existing network methods, based on functional connectivity. In summary, this paper presents a new computationally efficient and flexible method for directed network studies of the brain as well as new scientific findings regarding the functional organization of the human brain.

REFERENCES

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

BACCALÁ, L. A. and SAMESHIMA, K. (2001). Partial directed coherence: A new concept in neural structure determination. *Biol. Cybernet.* **84** 463–474.

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

BUCKNER, R. L., ANDREWS-HANNA, J. R. and SCHACTER, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Ann. N.Y. Acad. Sci.* **1124** 1–38. https://doi.org/10.1196/annals.1440.011

BUSHARA, K. O., GRAFMAN, J. and HALLETT, M. (2001). Neural correlates of auditory–visual stimulus onset asynchrony detection. *J. Neurosci.* **21** 300–304.

CALHOUN, V. D. and ADALI, T. (2012). Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery. *IEEE Rev. Biomed. Eng.* **5** 60–73. https://doi.org/10.1109/RBME.2012.2211076

CHAIKIN, P. M., LUBENSKY, T. C. and WITTEN, T. A. (1995). *Principles of Condensed Matter Physics* **10**.

CHIANG, S., GUINDANI, M., YEH, H. J., HANEEF, Z., STERN, J. M. and VANNUCCI, M. (2017). Bayesian vector autoregressive model for multi-subject effective connectivity inference using multi-modal neuroimaging data. *Hum. Brain Mapp.* **38** 1311–1332. https://doi.org/10.1002/hbm.23456

COLE, D. M., SMITH, S. M. and BECKMANN, C. F. (2010). Advances and pitfalls in the analysis and interpretation of resting-state FMRI data. *Front. Syst. Neurosci.* **4** 8. https://doi.org/10.3389/fnsys.2010.00008

Key words and phrases. Directed connectivity, brain network, variational Bayes, resting-state fMRI.

CRADDOCK, R. C., JAMES, G. A., HOLTZHEIMER III, P. E., HU, X. P. and MAYBERG, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum. Brain Mapp.* **33** 1914–1928.

DAVEY, C. G., PUJOL, J. and HARRISON, B. J. (2016). Mapping the self in the brain's default mode network. *NeuroImage* **132** 390–397. https://doi.org/10.1016/j.neuroimage.2016.02.022

DAVID, O. and FRISTON, K. (2003). A neural mass model for MEG/EEG: Coupling and neuronal dynamics. *NeuroImage* **20** 1743–1755.

DESHPANDE, G., HU, X., STILLA, R. and SATHIAN, K. (2008). Effective connectivity during haptic perception: A study using Granger causality analysis of functional magnetic resonance imaging data. *NeuroImage* **40** 1807–1814.

DE LUCA, M., BECKMANN, C. F., DE STEFANO, N., MATTHEWS, P. M. and SMITH, S. M. (2006). FMRI resting state networks define distinct modes of long-distance interactions in the human brain. *NeuroImage* **29** 1359–1367.

DURANTE, D. and DUNSON, D. B. (2014). Nonparametric Bayes dynamic modelling of relational data. *Biometrika* **101** 883–898. MR3286923 https://doi.org/10.1093/biomet/asu040

FIENBERG, S. E., MEYER, M. M. and WASSERMAN, S. S. (1985). Statistical analysis of multiple sociometric relations. *J. Amer. Statist. Assoc.* **80** 51–67.

FODOR, J. A. (1983). *The Modularity of Mind*. MIT Press, Cambridge.

FRÄSSLE, S., LOMAKINA, E. I., KASPER, L., MANJALY, Z. M., LEFF, A., PRUESSMANN, K. P., BUHMANN, J. M. and STEPHAN, K. E. (2018). A generative model of whole-brain effective connectivity. *NeuroImage* **179** 505–529. https://doi.org/10.1016/j.neuroimage.2018.05.058

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2014). Glasso: Graphical lasso-estimation of Gaussian graphical models. *R Package Version* **1**.

FRISTON, K. (1994). Functional and effective connectivity in neuroimaging: A synthesis. *Hum. Brain Mapp.* **2** 56–78.

FRISTON, K. (2011). Functional and effective connectivity: A review. *Brain Connect.* **1** 13–36.

FRISTON, K., HARRISON, L. and PENNY, W. (2003). Dynamic causal modelling. *NeuroImage* **19** 1273–1302.

GARGOURI, F., KALLEL, F., DELPHINE, S., HAMIDA, A. B., LEHÉRICY, S. and VALABREGUE, R. (2018). The influence of preprocessing steps on graph theory measures derived from resting state fMRI. *Front. Comput. Neurosci.* **12** 8. https://doi.org/10.3389/fncom.2018.00008

GLASSER, M. F., SOTIROPOULOS, S. N., WILSON, J. A., COALSON, T. S., FISCHL, B., ANDERSSON, J. L., XU, J., JBABDI, S., WEBSTER, M. et al. (2013). The minimal preprocessing pipelines for the human connectome project. *NeuroImage* **80** 105–124.

GUSNARD, D. A. and RAICHLE, M. E. (2001). Searching for a baseline: Functional imaging and the resting human brain. *Nat. Rev. Neurosci.* **2** 685–694.

HAYDEN, D., CHANG, Y. H., GONCALVES, J. and TOMLIN, C. J. (2016). Sparse network identifiability via compressed sensing. *Automatica J. IFAC* **68** 9–17. MR3483663 https://doi.org/10.1016/j.automatica.2016.01.008

HE, Y., WANG, J., WANG, L., CHEN, Z. J., YAN, C., YANG, H., TANG, H., ZHU, C., GONG, Q. et al. (2009). Uncovering intrinsic modular organization of spontaneous brain activity in humans. *PLoS ONE* **4** e5226.

HINRICHS, H., HEINZE, H. J. and SCHOENFELD, M. A. (2006). Causal visual interactions as revealed by an information theoretic measure and fMRI. *NeuroImage* **31** 1051–1060. https://doi.org/10.1016/j.neuroimage.2006.01.038

HOFFMAN, M. and BLEI, D. (2015). Stochastic structured variational inference. In *Artificial Intelligence and Statistics PMLR* 361–369.

HOLMES, E. E., WARD, E. J. and WILLS, K. (2012). MARSS: Multivariate autoregressive state-space models for analyzing time-series data. *R J.* **4**.

IKEDA, S., KAWANO, K., WATANABE, S., YAMASHITA, O. and KAWAHARA, Y. (2022). Predicting behavior through dynamic modes in resting-state fMRI data. *NeuroImage* **247** 118801. https://doi.org/10.1016/j.neuroimage.2021.118801

KONTOGHIORGHES, E. J., ed. (2006) *Handbook of Parallel Computing and Statistics*. *Statistics*: *Textbooks and Monographs* **184**. CRC Press/CRC, Boca Raton, FL. MR2265409

KOOK, J. H., VAUGHN, K. A., DEMASTER, D. M., EWING-COBBS, L. and VANNUCCI, M. (2020). BVAR-connect: A variational Bayes approach to multi-subject vector autoregressive models for inference on brain connectivity networks. *Neuroinformatics* 1–18.

KORZENIEWSKA, A., CERVENKA, M. C., JOUNY, C. C., PERILLA, J. R., HAREZLAK, J., BERGEY, G. K., FRANASZCZUK, P. J. and CRONE, N. E. (2014). Ictal propagation of high frequency activity is recapitulated in interictal recordings: Effective connectivity of epileptogenic networks recorded with intracranial EEG. *NeuroImage* **101** 96–113. https://doi.org/10.1016/j.neuroimage.2014.06.078

KRAMER, M., KOLACZYK, E. and KIRSCH, H. (2008). Emergent network topology at seizure onset in humans. *Epilepsy Res.* **79** 173–186.

LI, H., WANG, Y., YAN, G., SUN, Y., TANABE, S., LIU, C.-C., QUIGG, M. S. and ZHANG, T. (2021). A Bayesian state-space approach to mapping directional brain networks. *J. Amer. Statist. Assoc.* **116** 1637–1647. MR4353702 https://doi.org/10.1080/01621459.2020.1865985

LINDQUIST, M. A. (2008). The statistical analysis of fMRI data. *Statist. Sci.* **23** 439–464. MR2530545 https://doi.org/10.1214/09-STS282

LIU, Y. and AVIYENTE, S. (2012). Quantification of effective connectivity in the brain using a measure of directed information. *Comput. Math. Methods Med.* Art. ID 635103, 16. MR2926471 https://doi.org/10.1155/2012/635103

MEJIA, A. F., NEBEL, M. B., WANG, Y., CAFFO, B. S. and GUO, Y. (2020). Template independent component analysis: Targeted and reliable estimation of subject-level brain networks using big data population priors. *J. Amer. Statist. Assoc.* **115** 1151–1177. MR4143456 https://doi.org/10.1080/01621459.2019.1679638

MENNES, M., KELLY, C., ZUO, X.-N., DI MARTINO, A., BISWAL, B. B., CASTELLANOS, F. X. and MILHAM, M. P. (2010). Inter-individual differences in resting-state functional connectivity predict task-induced BOLD activity. *NeuroImage* **50** 1690–1701.

MESULAM, M.-M. (1998). From sensation to cognition. *Brain*: *A Journal of Neurology* **121** 1013–1052.

MEUNIER, D., LAMBIOTTE, R., FORNITO, A., ERSCHE, K. D. and BULLMORE, E. T. (2009). Hierarchical modularity in human brain functional networks. *Front. Neuroinform.* **3** 37. https://doi.org/10.3389/neuro.11.037.2009

MOUSSA, M. N., STEEN, M. R., LAURIENTI, P. J. and HAYASAKA, S. (2012). Consistency of network modules in resting-state FMRI connectome data. *PLoS ONE* **7** e44428. https://doi.org/10.1371/journal.pone.0044428

NEWMAN, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577–8696.

NICHOLSON, W. B., MATTESON, D. S. and BIEN, J. (2017). VARX-L: Structured regularization for large vector autoregressions with exogenous variables. *Int. J. Forecast.* **33** 627–651.

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

PARK, H.-J. and FRISTON, K. (2013). Structural and functional brain networks: From connections to cognition. *Science* **342** 1238411. https://doi.org/10.1126/science.1238411

PENNY, W. D., FRISTON, K. J., ASHBURNER, J. T., KIEBEL, S. J. and NICHOLS, T. E. (2011). *Statistical Parametric Mapping*: *The Analysis of Functional Brain Images*. Elsevier, Amsterdam.

POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M. et al. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.

RAICHLE, M. E. (2015). The brain's default mode network. *Annu. Rev. Neurosci.* **38** 433–447.

RAICHLE, M. E., MACLEOD, A. M., SNYDER, A. Z., POWERS, W. J., GUSNARD, D. A. and SHULMAN, G. L. (2001). A default mode of brain function. *Proc. Natl. Acad. Sci. USA* **98** 676–682.

RIEDL, V., UTZ, L., CASTRILLÓN, G., GRIMMER, T., RAUSCHECKER, J. P., PLONER, M., FRISTON, K. J., DRZEZGA, A. and SORG, C. (2016). Metabolic connectivity mapping reveals effective connectivity in the resting human brain. *Proc. Natl. Acad. Sci. USA* **113** 428–433.

ROSENTHAL, J. S. (2000). Parallel computing and Monte Carlo algorithms. *Far East J. Theor. Stat.* **4** 207–236. MR1801279

SABESAN, S., GOOD, L. B., TSAKALIS, K. S., SPANIAS, A., TREIMAN, D. M. and IASEMIDIS, L. D. (2009). Information flow and application to epileptogenic focus localization from intracranial EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **17** 244–253. https://doi.org/10.1109/TNSRE.2009.2023291

SATO, J. R., FUJITA, A., CARDOSO, E. F., THOMAZ, C. E., BRAMMER, M. J. and AMARO JR, E. (2010). Analyzing the connectivity between regions of interest: An approach based on cluster Granger causality for fMRI data analysis. *NeuroImage* **52** 1444–1455.

SCHIFF, S., SAUER, T., KUMAR, R. and WEINSTEIN, S. (2005). Neuronal spatiotemporal pattern discrimination: The dynamical evolution of seizures. *NeuroImage* **28** 1043–1055.

SCHREIBER, T. (2000). Measuring information transfer. *Phys. Rev. Lett.* **85** 461–464. https://doi.org/10.1103/PhysRevLett.85.461

SCHRÖDER, A. L. and OMBAO, H. (2019). FreSpeD: Frequency-specific change-point detection in epileptic seizure multi-channel EEG data. *J. Amer. Statist. Assoc.* **114** 115–128. MR3941242 https://doi.org/10.1080/01621459.2018.1476238

SEELEY, W. W. (2019). The salience network: A neural system for perceiving and responding to homeostatic demands. *J. Neurosci.* **39** 9878–9882.

SEELEY, W. W., MENON, V., SCHATZBERG, A. F., KELLER, J., GLOVER, G. H., KENNA, H., REISS, A. L. and GREICIUS, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *J. Neurosci.* **27** 2349–2356.

SMITH, S. M., BECKMANN, C. F., ANDERSSON, J., AUERBACH, E. J., BIJSTERBOSCH, J., DOUAUD, G., DUFF, E., FEINBERG, D. A., GRIFFANTI, L. et al. (2013). Resting-state fMRI in the human connectome project. *NeuroImage* **80** 144–168.

SMITH, S. M., FOX, P. T., MILLER, K. L., GLAHN, D. C., FOX, P. M., MACKAY, C. E., FILIPPINI, N., WATKINS, K. E., TORO, R. et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. USA* **106** 13040–13045.

SPORNS, O. (2011). *Networks of the Brain*. The MIT Press, Cambridge, MA.

SPORNS, O. (2013). Network attributes for segregation and integration in the human brain. *Curr. Opin. Neurobiol.* **23** 162–171.

SPORNS, O. and BETZEL, R. F. (2016). Modular brain networks. *Annu. Rev. Psychol.* **67** 613–640.

SPORNS, O., HONEY, C. J. and KÖTTER, R. (2007). Identification and classification of hubs in brain networks. *PLoS ONE* **2** e1049. https://doi.org/10.1371/journal.pone.0001049

VAN MIERLO, P., CARRETTE, E., HALLEZ, H., RAEDT, R., MEURS, A., VANDENBERGHE, S., ROOST, D. V., BOON, P., STAELENS, S. et al. (2013). Ictal-onset localization through connectivity analysis of intracranial EEG signals in patients with refractory epilepsy. *Epilepsia* **54** 1409–1418. https://doi.org/10.1111/epi.12206

VAN DE VEN, V. G., FORMISANO, E., PRVULOVIC, D., ROEDER, C. H. and LINDEN, D. E. (2004). Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. *Hum. Brain Mapp.* **22** 165–178.

VAN ESSEN, D. C., SMITH, S. M., BARCH, D. M., BEHRENS, T. E., YACOUB, E., UGURBIL, K., CONSORTIUM, W.-M. H. et al. (2013). The Wu-Minn human connectome project: An overview. *NeuroImage* **80** 62–79.

VICENTE, R., WIBRAL, M., LINDNER, M. and PIPA, G. (2011). Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *J. Comput. Neurosci.* **30** 45–67. MR2774389 https://doi.org/10.1007/s10827-010-0262-3

WAINWRIGHT, M. J. and JORDAN, M. I. (2008). *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers, Hanover.

WANG, Y., YAN, G., WANG, X., LI, S., PENG, L., TUDORASCU, D. L. and ZHANG, T. (2023). Supplement to "A variational Bayesian approach to identifying whole-brain directed Networks with fMRI data." https://doi.org/10.1214/22-AOAS1640SUPPA, https://doi.org/10.1214/22-AOAS1640SUPPB

WILKE, C., WORRELL, G. and HE, B. (2011). Graph analysis of epileptogenic networks in human partial epilepsy. *Epilepsia* **52** 84–93.

WITTEN, D. M., FRIEDMAN, J. H. and SIMON, N. (2011). New insights and faster computations for the graphical lasso. *J. Comput. Graph. Statist.* **20** 892–900. MR2878953 https://doi.org/10.1198/jcgs.2011.11051a

XIA, M., WANG, J. and HE, Y. (2013). BrainNet viewer: A network visualization tool for human brain connectomics. *PLoS ONE* **8** e68910.

YAN, C.-G., WANG, X.-D., ZUO, X.-N. and ZANG, Y.-F. (2016). DPABI: Data processing & analysis for (resting-state) brain imaging. *Neuroinformatics* **14** 339–351.

ZEKI, S., WATSON, J., LUECK, C., FRISTON, K. J., KENNARD, C. and FRACKOWIAK, R. (1991). A direct demonstration of functional specialization in human visual cortex. *J. Neurosci.* **11** 641–649.

ZHANG, T., SUN, Y., YAN, G., YIN, Q., LI, H., TANABE, S., CAFFO, B. and QUIGG, M. (2019). Bayesian inference of a directional brain network for intracranial EEG data. *Comput. Statist. Data Anal.* **106847**.

ZHANG, T., WU, J., LI, F., CAFFO, B. and BOATMAN-REICH, D. (2015). A dynamic directional model for effective brain connectivity using electrocorticographic (ECoG) time series. *J. Amer. Statist. Assoc.* **110** 93–106. MR3338489 https://doi.org/10.1080/01621459.2014.988213

ZHANG, T., YIN, Q., CAFFO, B., SUN, Y. and BOATMAN-REICH, D. (2017). Bayesian inference of high-dimensional, cluster-structured ordinary differential equation models with applications to brain connectivity studies. *Ann. Appl. Stat.* **11** 868–897. MR3693550 https://doi.org/10.1214/17-AOAS1021

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 539–559
https://doi.org/10.1214/22-AOAS1641

SUBJECT-SPECIFIC DIRICHLET-MULTINOMIAL REGRESSION FOR MULTI-DISTRICT MICROBIOTA DATA ANALYSIS

BY MATTEO PEDONE[1,a], AMEDEO AMEDEI[2,c] AND FRANCESCO C. STINGO[1,b]

[1]*Department of Statistics, Computer Science, Applications, University of Florence,* [a]*matteo.pedone@unifi.it,* [b]*francescoclaudio.stingo@unifi.it*

[2]*Department of Clinical and Experimental Medicine, University of Florence,* [c]*amedeo.amedei@unifi.it*

Many environments within the human body host a collection of microorganisms called microbiota. Recent findings have linked the composition of the microbiota to the development of different human diseases, including cancer. Motivated by a recent colorectal cancer (CRC) study, we investigate the effect of clinical factors and diet-related covariates on the microbiota compositions; for the patients enrolled in this study, microbiota abundance counts are collected from three different districts, namely, tumor, fecal and salivary samples. Building upon the Dirichlet-multinomial regression framework, we develop a high-dimensional Bayesian hierarchical model that exploits subject-specific regression coefficients to simultaneously borrow strength across districts and include complex interactions between diet and clinical factors if supported by the data. The proposed method identifies relevant associations through model selection priors and thresholding mechanisms. Posterior inference is performed via a Markov chain Monte Carlo algorithm. We use simulation studies to assess the performance of our method, and found our approach to outperform competing methods that do not account for complex interactions. Finally, a thorough analysis of the CRC data illustrates the benefits of the proposed approach.

REFERENCES

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192 https://doi.org/10.1214/009053604000000238

BESCOS, R., ASHWORTH, A., CUTLER, C., BROOKES, Z. L., BELFIELD, L., RODILES, A., CASAS-AGUSTENCH, P., FARNHAM, G., LIDDLE, L. et al. (2020). Effects of Chlorhexidine mouthwash on the oral microbiome. *Sci. Rep.* **10** 1–8.

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics* 73–80.

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751 https://doi.org/10.1093/biomet/asq017

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

CHIPMAN, H. (1996). Bayesian variable selection with related predictors. *Canad. J. Statist.* **24** 17–36. MR1394738 https://doi.org/10.2307/3315687

CHIPMAN, H., GEORGE, E. I. and MCCULLOCH, R. E. (2001). The practical implementation of Bayesian model selection. In *Model Selection*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **38** 65–134. IMS, Beachwood, OH. With discussion by M. Clyde, Dean P. Foster, and Robert A. Stine, and a rejoinder by the authors. MR2000752 https://doi.org/10.1214/lnms/1215540964

DE BONO, J. S. and ASHWORTH, A. (2010). Translating cancer research into targeted therapeutics. *Nature* **467** 543–549.

EDDELBUETTEL, D. and SANDERSON, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Comput. Statist. Data Anal.* **71** 1054–1063. MR3132026 https://doi.org/10.1016/j.csda.2013.02.005

EDDELBUETTEL, D., FRANÇOIS, R., ALLAIRE, J., USHEY, K., KOU, Q., RUSSEL, N., CHAMBERS, J. and BATES, D. (2011). Rcpp: Seamless R and C++ integration. *J. Stat. Softw.* **40** 1–18.

Key words and phrases. Microbiota data analysis, subject-specific coefficients, Bayesian hierarchical models, multivariate count data, model selection.

FAITH, J. J., GURUGE, J. L., CHARBONNEAU, M., SUBRAMANIAN, S., SEEDORF, H., GOODMAN, A. L., CLEMENTE, J. C., KNIGHT, R., HEATH, A. C. et al. (2013). The long-term stability of the human gut microbiota. *Science* **341**.

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal*. **1** 515–533. MR2221284 https://doi.org/10.1214/06-BA117A

GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat*. **2** 1360–1383. MR2655663 https://doi.org/10.1214/08-AOAS191

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

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

GRIFFIN, J. and BROWN, P. (2017). Hierarchical shrinkage priors for regression models. *Bayesian Anal*. **12** 135–159. MR3597570 https://doi.org/10.1214/15-BA990

GUSTAFSON, P. (2000). Bayesian regression modeling with interactions and smooth effects. *J. Amer. Statist. Assoc*. **95** 795–806.

HARO, C., RANGEL-ZUNIGA, O. A., ALCALÀ-DIAZ, J. F., GÒMEZ-DELGADO, F., PÈREZ-MARTINEZ, P., DELGADO-LISTA, J., QUINTANA-NAVARRO, G. M., LANDA, B. B., NAVAS-CORTÈS, J. A. et al. (2016). Intestinal microbiota is influenced by gender and body mass index. *PLoS ONE* **11** e0154090.

HARRISON, J. G., CALDER, W. J., SHASTRY, V. and BUERKLE, C. A. (2020). Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Mol. Ecol. Resour*. **20** 481–497. https://doi.org/10.1111/1755-0998.13128

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. With discussion and a reply by the authors. MR1229881

HENTGES, D. J., MAIER, B. R., BURTON, G. C., FLYNN, M. A. and TSUTAKAWA, R. K. (1977). Effect of a high-beef diet on the fecal bacterial flora of humans. *Cancer Res*. **37** 568–571.

JAHANI-SHERAFAT, S., ALEBOUYEH, M., MOGHIM, S., AMOLI, H. A. and GHASEMIAN-SAFAEI, H. (2018). Role of gut microbiota in the pathogenesis of colorectal cancer; a review article. *Gastroenterology and Hepatology from Bed to Bench* **11** 101.

JENKINS, T. P., FORMENTI, F., CASTRO, C., PIUBELLI, C., PERANDIN, F., BUONFRATE, D., OTRANTO, D., GRIFFIN, J. L., KRAUSE, L. et al. (2018). A comprehensive analysis of the faecal microbiome and metabolome of Strongyloides stercoralis infected volunteers from a non-endemic area. *Sci. Rep*. **8** 1–13.

KOSLOVSKY, M. D., HOFFMAN, K. L., DANIEL, C. R. and VANNUCCI, M. (2020). A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. *Ann. Appl. Stat*. **14** 1471–1492. MR4152142 https://doi.org/10.1214/20-AOAS1354

KOSOROK, M. R. and LABER, E. B. (2019). Precision medicine. *Annu. Rev. Stat. Appl*. **6** 263–286. MR3939521 https://doi.org/10.1146/annurev-statistics-030718-105251

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

LI, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl*. **2** 73–94.

LIN, W., SHI, P., FENG, R. and LI, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101** 785–797. MR3286917 https://doi.org/10.1093/biomet/asu031

MANDRIOLI, J., AMEDEI, A., CAMMAROTA, G., NICCOLAI, E., ZUCCHI, E., D'AMICO, R., RICCI, F., QUARANTA, G., SPANU, T. et al. (2019). FETR-ALS study protocol: A randomized clinical trial of fecal microbiota transplantation in amyotrophic lateral sclerosis. *Front. Neurol*. **10** 1021. https://doi.org/10.3389/fneur.2019.01021

MAO, J. and MA, L. (2020). Dirichlet-tree multinomial mixtures for clustering microbiome compositions. arXiv preprint arXiv:2008.00400.

MARTIN, I., UH, H.-W., SUPALI, T., MITREVA, M. and HOUWING-DUISTERMAAT, J. J. (2019). The mixed model for the analysis of a repeated-measurement multivariate count data. *Stat. Med*. **38** 2248–2268. MR3939815 https://doi.org/10.1002/sim.8101

MATSUOKA, K. and KANAI, T. (2015). The gut microbiota and inflammatory bowel disease. *Springer Semin. Immunopathol*. **37** 47–55.

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc*. **83** 1023–1036. With comments by James Berger and C. L. Mallows and with a reply by the authors. MR0997578

MOSIMANN, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* **49** 65–82. MR0143299 https://doi.org/10.1093/biomet/49.1-2.65

NEAL, R. M. (2003). Slice sampling. *Ann. Statist.* **31** 705–767. With discussions and a rejoinder by the author. MR1994729 https://doi.org/10.1214/aos/1056562461

NI, Y., STINGO, F. C. and BALADANDAYUTHAPANI, V. (2019). Bayesian graphical regression. *J. Amer. Statist. Assoc.* **114** 184–197. MR3941247 https://doi.org/10.1080/01621459.2017.1389739

NI, Y., STINGO, F. C., HA, M. J., AKBANI, R. and BALADANDAYUTHAPANI, V. (2019). Bayesian hierarchical varying-sparsity regression models with application to cancer proteogenomics. *J. Amer. Statist. Assoc.* **114** 48–60. MR3941237 https://doi.org/10.1080/01621459.2018.1434529

NICCOLAI, E., RUSSO, E., BALDI, S., RICCI, F., NANNINI, G., PEDONE, M., STINGO, F. C., TADDEI, A., RINGRESSI, M. N. et al. (2020). Significant and conflicting correlation of IL-9 with Prevotella and Bacteroides in human colorectal cancer. *Front. Immunol.* **11**.

PEDONE, M., AMEDEI, A. and STINGO, F. C. (2023). Supplement to "Subject-specific Dirichlet-multinomial regression for multi-district microbiota data analysis." https://doi.org/10.1214/22-AOAS1641SUPPA, https://doi.org/10.1214/22-AOAS1641SUPPB

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2014). The Bayesian bridge. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 713–733. MR3248673 https://doi.org/10.1111/rssb.12042

REN, B., BACALLADO, S., FAVARO, S., VATANEN, T., HUTTENHOWER, C. and TRIPPA, L. (2020). Bayesian mixed effects models for zero-inflated compositions in microbiome data analysis. *Ann. Appl. Stat.* **14** 494–517. MR4085103 https://doi.org/10.1214/19-AOAS1295

ROSSELL, D. and TELESCA, D. (2017). Nonlocal priors for high-dimensional estimation. *J. Amer. Statist. Assoc.* **112** 254–265. MR3646569 https://doi.org/10.1080/01621459.2015.1130634

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720 https://doi.org/10.1017/CBO9780511755453

RUSSO, E., BACCI, G., CHIELLINI, C., FAGORZI, C., NICCOLAI, E., TADDEI, A., RICCI, F., RINGRESSI, M. N., BORRELLI, R. et al. (2018). Preliminary comparison of oral and intestinal human microbiota in patients with colorectal cancer: A pilot study. *Front. Microbiol.* **8** 2699.

SANZ, Y., OLIVARES, M., MOYA-PÉREZ, Á. and AGOSTONI, C. (2015). Understanding the role of gut microbiome in metabolic disease risk. *Pediatr. Res.* **77** 236–244.

SAVITSKY, T., VANNUCCI, M. and SHA, N. (2011). Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statist. Sci.* **26** 130–149. MR2849913 https://doi.org/10.1214/11-STS354

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

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450 https://doi.org/10.1214/10-AOS792

SHULER, K., SISON-MANGUS, M. and LEE, J. (2020). Bayesian sparse multivariate regression with asymmetric nonlocal priors for microbiome data analysis. *Bayesian Anal.* **15** 559–578. MR4078725 https://doi.org/10.1214/19-BA1164

STINGO, F. C., CHEN, Y. A., TADESSE, M. G. and VANNUCCI, M. (2011). Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Ann. Appl. Stat.* **5** 1978–2002. MR2884929 https://doi.org/10.1214/11-AOAS463

TAI, N., WONG, F. S. and WEN, L. (2015). The role of gut microbiota in the development of type 1, type 2 diabetes mellitus and obesity. *Rev. Endocr. Metab. Disord.* **16** 55–65.

TANG, Y., MA, L. and NICOLAE, D. L. (2018). A phylogenetic scan test on a Dirichlet-tree multinomial model for microbiome data. *Ann. Appl. Stat.* **12** 1–26. MR3773384 https://doi.org/10.1214/17-AOAS1086

TSAGRIS, M., ALENAZI, A. and STEWART, C. (2021). Non-parametric regression models for compositional data.

TURNBAUGH, P. J., LEY, R. E., HAMADY, M., FRASER-LIGGETT, C. M., KNIGHT, R. and GORDON, J. I. (2007). The human microbiome project. *Nature* **449** 804–810.

VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12** 1221–1274. With a rejoinder by the authors. MR3724985 https://doi.org/10.1214/17-BA1065

VIEIRA, S. M., PAGOVICH, O. E. and KRIEGEL, M. A. (2014). Diet, microbiota and autoimmune diseases. *Lupus* **23** 518–526.

WADSWORTH, W. D., ARGIENTO, R., GUINDANI, M., GALLOWAY-PENA, J., SHELBURNE, S. A. and VANNUCCI, M. (2017). An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinform.* **18** 1–12.

XUN, Z., ZHANG, Q., XU, T., CHEN, N. and CHEN, F. (2018). Dysbiosis and ecotypes of the salivary microbiome associated with inflammatory bowel diseases and the assistance in diagnosis of diseases using oral bacterial profiles. *Front. Microbiol.* **9** 1136.

ZHANG, H. and SUN, L. (2018). When human cells meet bacteria: Precision medicine for cancers using the microbiota. *Am. J. Cancer Res.* **8** 1157.

ZHANG, Y., ZHOU, H., ZHOU, J. and SUN, W. (2017). Regression models for multivariate count data. *J. Comput. Graph. Statist.* **26** 1–13. MR3610402 https://doi.org/10.1080/10618600.2016.1154063

ZHANG, L., SHI, Y., JENQ, R. R., DO, K.-A. and PETERSON, C. B. (2021). Bayesian compositional regression with structured priors for microbiome feature selection. *Biometrics* **77** 824–838. MR4320660 https://doi.org/10.1111/biom.13335

ZHU, Y., LIN, X., LI, H., LI, Y., SHI, X., ZHAO, F., XU, X., LI, C. and ZHOU, G. (2016). Intake of meat proteins substantially increased the relative abundance of genus Lactobacillus in rat feces. *PLoS ONE* **11** e0152678.

ZHU, Y., LIN, X., ZHAO, F., SHI, X., LI, H., LI, Y., ZHU, W., XU, X., LI, C. et al. (2015). Meat, dairy and plant proteins alter bacterial composition of rat gut bacteria. *Sci. Rep.* **5** 1–14.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 560–582
https://doi.org/10.1214/22-AOAS1642

SPATIOTEMPORAL WILDFIRE MODELING THROUGH POINT PROCESSES WITH MODERATE AND EXTREME MARKS

BY JONATHAN KOH[1,2,a], FRANÇOIS PIMONT[3,b], JEAN-LUC DUPUY[3,c] AND THOMAS OPITZ[4,d]

[1] *Institute of Mathematics, EPFL*

[2] *Institute of Mathematical Statistics and Actuarial Science, Oeschger Centre for Climate Change Research, University of Bern,* [a] *jonathan.koh@unibe.ch*

[3] *URFM UR629, INRAE,* [b] *francois.pimont@inrae.fr,* [c] *jean-luc.dupuy@inrae.fr*

[4] *BioSP UR546, INRAE,* [d] *thomas.opitz@inrae.fr*

Accurate spatiotemporal modeling of conditions leading to moderate and large wildfires provides better understanding of mechanisms driving fire-prone ecosystems and improves risk management. Here, we develop a joint model for the occurrence intensity and the wildfire size distribution, by combining extreme-value theory and point processes within a novel Bayesian hierarchical model, and use it to study daily summer wildfire data for the French Mediterranean basin during 1995–2018. The occurrence component models wildfire ignitions as a spatiotemporal log-Gaussian Cox process. Burnt areas are numerical marks attached to points and are considered as extreme if they exceed a high threshold. The size component is a two-component mixture varying in space and time that jointly models moderate and extreme fires. We capture nonlinear influence of covariates (Fire Weather Index, forest cover) through component-specific smooth functions which may vary with season. We propose estimating shared random effects between model components to reveal and interpret common drivers of different aspects of wildfire activity. This increases parsimony and reduces estimation uncertainty, giving better predictions. Specific stratified subsampling of zero counts is implemented to cope with large observation vectors. We compare and validate models through predictive scores and visual diagnostics. Our methodology provides a holistic approach to explaining and predicting the drivers of wildfire activity and associated uncertainties.

REFERENCES

BADDELEY, A., BERMAN, M., FISHER, N. I., HARDEGEN, A., MILNE, R. K., SCHUHMACHER, D., SHAH, R. and TURNER, R. (2010). Spatial logistic regression and change-of-support in Poisson point processes. *Electron. J. Stat.* **4** 1151–1201. MR2735883 https://doi.org/10.1214/10-EJS581

BADDELEY, A., COEURJOLLY, J.-F., RUBAK, E. and WAAGEPETERSEN, R. (2014). Logistic regression for spatial Gibbs point processes. *Biometrika* **101** 377–392. MR3215354 https://doi.org/10.1093/biomet/ast060

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

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

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

BOLIN, D. and LINDGREN, F. (2015). Excursion and contour uncertainty regions for latent Gaussian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 85–106. MR3299400 https://doi.org/10.1111/rssb.12055

BOLIN, D. and WALLIN, J. (2020). Scale dependence: Why the average CRPS often is inappropriate for ranking probabilistic forecasts. ArXiv Preprint. Available at arXiv:1912.05642.

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

Key words and phrases. Bayesian hierarchical model, Cox process, extreme-value theory, forest fires, shared random effects.

COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. *Springer Series in Statistics*. Springer, London. MR1932132 https://doi.org/10.1007/978-1-4471-3675-0

CUI, W. and PERERA, A. H. (2008). What do we know about forest fire size distribution, and why is this knowledge useful for forest management? *Int. J. Wildland Fire* **17** 234–244.

CUMMING, S. (2001). A parametric model of the fire-size distribution. *Can. J. For. Res.* **31** 1297–1303.

DAVISON, A. C. and SMITH, R. L. (1990). Models for exceedances over high thresholds. *J. Roy. Statist. Soc. Ser. B* **52** 393–442. MR1086795

DE ZEA BERMUDEZ, P., MENDES, J., PEREIRA, J. M. C., TURKMAN, K. F. and VASCONCELOS, M. J. P. (2009). Spatial and temporal extremes of wildfire sizes in Portugal (1984–2004). *Int. J. Wildland Fire* **18** 983–991.

EVIN, G., CURT, T. and ECKERT, N. (2018). Has fire policy decreased the return period of the largest wildfire events in France? A Bayesian assessment based on extreme value theory. *Nat. Hazards Earth Syst. Sci.* **18** 2641–2651.

FAWCETT, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.* **27** 861–874.

FUGLSTAD, G.-A., SIMPSON, D., LINDGREN, F. and RUE, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *J. Amer. Statist. Assoc.* **114** 445–452. MR3941267 https://doi.org/10.1080/01621459.2017.1415907

GENTON, M. G., BUTRY, D. T., GUMPERTZ, M. L. and PRESTEMON, J. P. (2006). Spatio-temporal analysis of wildfire ignitions in the St Johns River water management district, Florida. *Int. J. Wildland Fire* **15** 87–97.

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

ILLIAN, J. B., SØRBYE, S. H. and RUE, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *Ann. Appl. Stat.* **6** 1499–1530. MR3058673 https://doi.org/10.1214/11-AOAS530

JONES, M. W., SMITH, A., BETTS, R., CANADELL, J. G., PRENTICE, I. C. and LE QUÉRÉ, C. (2020). ScienceBrief review: Climate change increases the risk of wildfires. In *Critical Issues in Climate Change Science* (C. Le Quéré, P. Liss and P. Forster, eds.).

JOSEPH, M. B., ROSSI, M. W., MIETKIEWICZ, N. P., MAHOOD, A. L., CATTAU, M. E., ST. DENIS, L. A., NAGY, R. C., IGLESIAS, V., ABATZOGLOU, J. T. et al. (2019). Spatiotemporal prediction of wildfire size extremes with Bayesian finite sample maxima. *Ecol. Appl.* **29** e01898.

KOH, J., PIMONT, F., DUPUY, J.-L and OPITZ, T. (2023). Supplement to "Spatiotemporal wildfire modeling through point processes with moderate and extreme marks." https://doi.org/10.1214/22-AOAS1642SUPP

KRAINSKI, E. T., GÓMEZ-RUBIO, V., BAKKA, H., LENZI, A., CASTRO-CAMILO, D., SIMPSON, D., LINDGREN, F. and RUE, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. CRC Press/CRC, Boca Raton.

LINDGREN, F. and RUE, H. (2015). Bayesian spatial modelling with R-INLA. *J. Stat. Softw.* **63**.

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

MENDES, J. M., DE ZEA BERMUDEZ, P. C., PEREIRA, J., TURKMAN, K. F. and VASCONCELOS, M. J. P. (2010). Spatial extremes of wildfire sizes: Bayesian hierarchical models for extremes. *Environ. Ecol. Stat.* **17** 1–28. MR2594933 https://doi.org/10.1007/s10651-008-0099-3

MORITZ, M. A. (1997). Analyzing extreme disturbance events: Fire in los padres national forest. *Ecol. Appl.* **7** 1252–1262.

NORTHROP, P. J. and COLEMAN, C. L. (2014). Improved threshold diagnostic plots for extreme value analyses. *Extremes* **17** 289–303. MR3207718 https://doi.org/10.1007/s10687-014-0183-z

OPITZ, T. (2017). Latent Gaussian modeling and INLA: A review with focus on space-time applications. *J. SFdS* **158** 62–85. MR3720130

OPITZ, T., BAKKA, H., HUSER, R. and LOMBARDO, L. (2022). High-resolution Bayesian mapping of landslide hazard with unobserved trigger event. *Ann. Appl. Stat.* **16** 1653–1675. https://doi.org/10.1214/21-AOAS1561

OPITZ, T., BONNEU, F. and GABRIEL, E. (2020). Point-process based Bayesian modeling of space-time structures of forest fire occurrences in Mediterranean France. *Spat. Stat.* **40** 100429. MR4181143 https://doi.org/10.1016/j.spasta.2020.100429

OPITZ, T., HUSER, R., BAKKA, H. and RUE, H. (2018). INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. *Extremes* **21** 441–462. MR3855716 https://doi.org/10.1007/s10687-018-0324-x

PENG, R. D., SCHOENBERG, F. P. and WOODS, J. A. (2005). A space-time conditional intensity model for evaluating a wildfire hazard index. *J. Amer. Statist. Assoc.* **100** 26–35. MR2166067 https://doi.org/10.1198/016214504000001763

PEREIRA, J. M. C. and TURKMAN, K. F. (2019). Statistical models of vegetation fires: Spatial and temporal patterns. In *Handbook of Environmental and Ecological Statistics*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 401–420. CRC Press, Boca Raton, FL. MR3889904

PIMONT, F., FARGEON, H., OPITZ, T., RUFFAULT, J., BARBERO, R., MARTIN-STPAUL, N., RIGOLOT, E. I., RIVIÈRE, M. and DUPUY, J.-L. (2021). Prediction of regional wildfire activity in the probabilistic Bayesian framework of firelihood. *Ecol*. *Appl*. e02316.

PREISLER, H. K., BRILLINGER, D. R., BURGAN, R. E. and BENOIT, J. (2004). Probability based models for estimation of wildfire risk. *Int*. *J*. *Wildland Fire* **13** 133–142.

RATHBUN, S. L. (2013). Optimal estimation of Poisson intensity with partially observed covariates. *Biometrika* **100** 277–281. MR3034343 https://doi.org/10.1093/biomet/ass069

RATHBUN, S. L., SHIFFMAN, S. and GWALTNEY, C. J. (2007). Modelling the effects of partially observed covariates on Poisson process intensity. *Biometrika* **94** 153–165. MR2367829 https://doi.org/10.1093/biomet/asm009

RÍOS-PENA, L., KNEIB, T., CADARSO-SUÁREZ, C., KLEIN, N. and MAREY-PÉREZ, M. (2018). Studying the occurrence and burnt area of wildfires using zero-one-inflated structured additive beta regression. *Environ*. *Model*. *Softw*. **110** 107–118.

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

RUE, H., RIEBLER, A., SØRBYE, S. H., ILLIAN, J. B., SIMPSON, D. P. and LINDGREN, F. K. (2017). Bayesian computing with INLA: A review. *Annu*. *Rev*. *Stat*. *Appl*. **4** 395–421.

RUFFAULT, J., MARTIN-STPAUL, N., PIMONT, F. and DUPUY, J.-L. (2018). How well do meteorological drought indices predict live fuel moisture content (LFMC)? An assessment for wildfire research and operations in Mediterranean ecosystems. *Agric*. *For*. *Meteorol*. **262** 391–401.

SCHOENBERG, F. P., PENG, R. and WOODS, J. (2003). On the distribution of wildfire sizes. *Environmetrics* **14** 583–592.

SERRA, L., JUAN, P., VARGA, D., MATEU, J. and SAEZ, M. (2013). Spatial pattern modelling of wildfires in Catalonia, Spain 2004–2008. *Environ*. *Model*. *Softw*. **40** 235–244.

SERRA, L., SAEZ, M., JUAN, P., VARGA, D. and MATEU, J. (2014). A spatio-temporal Poisson hurdle point process to model wildfires. *Stoch*. *Environ*. *Res*. *Risk Assess*. **28** 1671–1684.

SIMPSON, D., RUE, H., RIEBLER, A., MARTINS, T. G. and SØRBYE, S. H. (2017b). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist*. *Sci*. **32** 1–28. MR3634300 https://doi.org/10.1214/16-STS576

STEWART, S. I., RADELOFF, V. C., HAMMER, R. B. and HAWBAKER, T. J. (2007). Defining the wildland–urban interface. *J*. *For*. **105** 201–207.

TAYLOR, B. M. and DIGGLE, P. J. (2014). INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *J*. *Stat*. *Comput*. *Simul*. **84** 2266–2284. MR3223624 https://doi.org/10.1080/00949655.2013.788653

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J*. *Amer*. *Statist*. *Assoc*. **81** 82–86. MR0830567

TOKDAR, S. T. and KASS, R. E. (2010). Importance sampling: A review. *Wiley Interdiscip*. *Rev*.: *Comput*. *Stat*. **2** 54–60.

TONINI, M., PEREIRA, M. G., PARENTE, J. and OROZCO, C. V. (2017). Evolution of forest fires in Portugal: From spatio-temporal point events to smoothed density maps. *Nat*. *Hazards* **85** 1489–1510.

TURKMAN, K. F., TURKMAN, M. A. A. and PEREIRA, J. M. (2010). Asymptotic models and inference for extremes of spatio-temporal data. *Extremes* **13** 375–397. MR2733939 https://doi.org/10.1007/s10687-009-0092-8

VAN NIEKERK, J., BAKKA, H., RUE, H. and SCHENK, L. (2019). New frontiers in Bayesian modeling using the INLA package in R. ArXiv Preprint. Available at arXiv:1907.10426.

VAN WAGNER, C. E. (1977). Conditions for the start and spread of crown fire. *Can*. *J*. *For*. *Res*. **7** 23–34.

WANG, X., WOTTON, B. M., CANTIN, A. S., PARISIEN, M.-A., ANDERSON, K., MOORE, B. and FLANNIGAN, M. D. (2017). cffdrs: An R package for the Canadian forest fire danger rating system. *Ecol*. *Process*. **6** 5.

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

WOOLFORD, D. G., MARTELL, D. L., MCFAYDEN, C. B., EVENS, J., STACEY, A., WOTTON, B. M. and BOYCHUK, D. (2021). The development and implementation of a human-caused wildland fire occurrence prediction system for the province of Ontario, Canada. *Can*. *J*. *For*. *Res*. **51** 303–325.

Xi, D. D. Z., Taylor, S. W., Woolford, D. G. and Dean, C. B. (2019). Statistical models of key components of wildfire risk. *Annu*. *Rev*. *Stat*. *Appl*. **6** 197–222. MR3939518 https://doi.org/10.1146/annurev-statistics-031017-100450

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

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 583–605
https://doi.org/10.1214/22-AOAS1643

BAYESIAN CLUSTERING OF SPATIAL FUNCTIONAL DATA WITH APPLICATION TO A HUMAN MOBILITY STUDY DURING COVID-19

BY BOHAI ZHANG[1,a], HUIYAN SANG[2,b], ZHAO TANG LUO[2,c] AND HUI HUANG[3,d]

[1]*School of Statistics and Data Science, LPMC, and KLMDASR, Nankai University,* [a]*bohaizhang@nankai.edu.cn*
[2]*Department of Statistics, Texas A&M University,* [b]*huiyan@stat.tamu.edu,* [c]*zt.luo@tamu.edu*
[3]*School of Mathematics, Sun Yat-sen University,* [d]*huangh89@mail.sysu.edu.cn*

The coronavirus (COVID-19) global pandemic has made a significant impact on people's social activities. Cell phone mobility data provide unique and rich information on studying this impact. The motivating dataset of this study is the daily leaving-home index data at Harris County in Texas provided by SafeGraph. To study changes in daily leaving-home index and how they relate to public policy and sociodemographic variables, we propose a new Bayesian wavelet model for modeling and clustering spatial functional data, where domain partitioning is achieved by operating on the spanning trees. The resulting clusters can have arbitrary shapes and are spatially contiguous in the input domain. An efficient tailored reversible jump Markov chain Monte Carlo algorithm is proposed to implement the model. The method is applied to the spatial functional data of the daily percentages of people who left home. We focus on the time period covering both lockdown and phased reopening in Texas during the COVID-19 pandemic and study the changing behaviors of those functional curves. By linking the clustering results with the sociodemographic information, we identify several covariates of census blocks that have a noticeable impact on the clustering patterns of people's mobility behaviors.

REFERENCES

ABRAMOVICH, F., SAPATINAS, T. and SILVERMAN, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 725–749. MR1649547 https://doi.org/10.1111/1467-9868.00151

BANFIELD, J. D. and RAFTERY, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803–821. MR1243494 https://doi.org/10.2307/2532201

CHIOU, J.-M. and LI, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 679–699. MR2370075 https://doi.org/10.1111/j.1467-9868.2007.00605.x

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

CLYDE, M. and GEORGE, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 681–698. MR1796285 https://doi.org/10.1111/1467-9868.00257

COUTURE, V., DINGEL, J., GREEN, A., HANDBURY, J. and WILLIAMS, K. (2020). Exposure indices derived from PlaceIQ movement data. Available at https://github.com/COVIDExposureIndices/COVIDExposureIndices. Accessed 31 March 2022.

CSÁRDI, G. (2013). R Package igraph. Available at https://igraph.org/r/. Last accessed: 19/05/2022.

DAHL, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics* (K.-A. Do, P. Müller and M. Vannucci, eds.) **4** 201–218. Cambridge University Press.

DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **61**. SIAM, Philadelphia, PA. MR1162107 https://doi.org/10.1137/1.9781611970104

DENISON, D. G. T. and HOLMES, C. C. (2001). Bayesian partitioning for estimating disease risk. *Biometrics* **57** 143–149. MR1833300 https://doi.org/10.1111/j.0006-341X.2001.00143.x

Key words and phrases. Bayesian wavelet model, mobility behavior, reversible jump Markov chain Monte Carlo, spanning tree, spatial functional data.

DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika* **85** 363–377. MR1649118 https://doi.org/10.1093/biomet/85.2.363

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

DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464

ERWAN, G. F. (2019). R Package funHDDC-wavelet. Available at https://github.com/Erwangf/funHDDC-wavelet. Last accessed: 19/05/2022.

FRALEY, C. and RAFTERY, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *J. Classification* **24** 155–181. MR2415725 https://doi.org/10.1007/s00357-007-0004-5

GAO, S., RAO, J., KANG, Y., LIANG, Y., KRUSE, J., DOEPFER, D., SETHI, A. K., REYES, J. F. M., PATZ, J. (2020). Association of mobile phone location data indications of travel and stay-at-home mandates with COVID-19 infection rates in the US. *JAMA Netw. Open* **3** e2020485.

GIRALDO, R., DELICADO, P. and MATEU, J. (2012). Hierarchical clustering of spatially correlated functional data. *Stat. Neerl.* **66** 403–421. MR2983302 https://doi.org/10.1111/j.1467-9574.2012.00522.x

GUAN, W.-J., NI, Z.-Y., HU, Y., LIANG, W.-H., OU, C.-Q., HE, J.-X., LIU, L., SHAN, H., LEI, C.-L. et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* **382** 1708–1720.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810 https://doi.org/10.1093/biomet/82.4.711

GUAN, Y. and STEPHENS, M. (2011). Bayesian variable selection regression for Genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5** 1780–1815. MR2884922 https://doi.org/10.1214/11-AOAS455

HOLMES, C. C., DENISON, D. G. T., RAY, S. and MALLICK, B. K. (2005). Bayesian prediction via partitioning. *J. Comput. Graph. Statist.* **14** 811–830. MR2211368 https://doi.org/10.1198/106186005X78107

HSIANG, S., ALLEN, D., ANNAN-PHAN, S., BELL, K., BOLLIGER, I., CHONG, T., DRUCKENMILLER, H., HUANG, L. Y., HULTGREN, A. et al. (2020). The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* **584** 262–267.

HU, G., GENG, J., XUE, Y. and SANG, H. (2022). Bayesian spatial homogeneity pursuit of functional data: An application to the US income distribution. *Bayesian Anal.* https://doi.org/10.1214/22-BA1320

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

HYNDMAN, R. J. and SHANG, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *J. Comput. Graph. Statist.* **19** 29–45. MR2752026 https://doi.org/10.1198/jcgs.2009.08158

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

JAY, J., BOR, J., NSOESIE, E. O., LIPSON, S. K., JONES, D. K., GALEA, S. and RAIFMAN, J. (2020). Neighbourhood income and physical distancing during the COVID-19 pandemic in the United States. *Nat. Hum. Behav.* **4** 1294–1302.

JIANG, F., DENG, L., ZHANG, L., CAI, Y., CHEUNG, C. W. and XIA, Z. (2020). Review of the clinical characteristics of coronavirus disease 2019 (COVID-19). *J. Gen. Intern. Med.* **35** 1545–1549.

JIANG, H. and SERBAN, N. (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics* **54** 108–119. MR2929427 https://doi.org/10.1080/00401706.2012.657106

KIM, H.-M., MALLICK, B. K. and HOLMES, C. C. (2005). Analyzing nonstationary spatial data using piecewise Gaussian processes. *J. Amer. Statist. Assoc.* **100** 653–668. MR2160567 https://doi.org/10.1198/016214504000002014

KNORRHELD, L. and RASER, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56** 13–21.

KONOMI, B. A., SANG, H. and MALLICK, B. K. (2014). Adaptive Bayesian nonstationary modeling for large spatial datasets using covariance approximations. *J. Comput. Graph. Statist.* **23** 802–829. MR3224657 https://doi.org/10.1080/10618600.2013.812872

LEE, D. T. and SCHACHTER, B. J. (1980). Two algorithms for constructing a Delaunay triangulation. *Int. J. Comput. Inform. Sci.* **9** 219–242. MR0585950 https://doi.org/10.1007/BF00977785

LEWNARD, J. A. and LO, N. C. (2020). Scientific and ethical basis for social-distancing interventions against COVID-19. *Lancet Infect. Dis.* **20** 631–633. https://doi.org/10.1016/S1473-3099(20)30190-0

LI, F. and SANG, H. (2019). Spatial homogeneity pursuit of regression coefficients for large datasets. *J. Amer. Statist. Assoc.* **114** 1050–1062. MR4011757 https://doi.org/10.1080/01621459.2018.1529595

LUO, Z. T., SANG, H. and MALLICK, B. (2021). A Bayesian contiguous partitioning method for learning clustered latent variables. *J. Mach. Learn. Res.* **22** Paper No. 37, 52 pp. Available at https://jmlr.org/papers/v22/20-136.html. MR4253730

NASON, G. (2016). R package wavethresh. Available at https://cran.r-project.org/web/packages/wavethresh/wavethresh.pdf. Last accessed: 19/05/2022.

PAYNE, R. D., GUHA, N., DING, Y. and MALLICK, B. K. (2020). A conditional density estimation partition model using logistic Gaussian processes. *Biometrika* **107** 173–190. MR4064147 https://doi.org/10.1093/biomet/asz064

RAY, S. and MALLICK, B. (2006). Functional clustering by Bayesian wavelet methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 305–332. MR2188987 https://doi.org/10.1111/j.1467-9868.2006.00545.x

ROMANO, E., VERDE, R. and COZZA, V. (2011). Clustering spatial functional data: A method based on a nonparametric variogram estimation. In *New Perspectives in Statistical Modeling and Data Analysis. Stud. Classification Data Anal. Knowledge Organ.* 339–346. Springer, Heidelberg. MR3051225 https://doi.org/10.1007/978-3-642-11363-5_38

ROUSSEEUW, P. J., RUTS, I. and TUKEY, J. W. (1999). The bagplot: A bivariate boxplot. *Amer. Statist.* **53** 382–387.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SHENSA, M. J. (1992). The discrete wavelet transform: Wedding the a trous and Mallat algorithms. *IEEE Trans. Signal Process.* **40** 2464–2482.

SHENTAL, N., HERTZ, T., WEINSHALL, D. and PAVEL, M. (2002). Adjustment learning and relevant component analysis. In *European Conference on Computer Vision* 776–790. Springer, Berlin.

SUAREZ, A. J. and GHOSAL, S. (2016). Bayesian clustering of functional data using local features. *Bayesian Anal.* **11** 71–98. MR3447092 https://doi.org/10.1214/14-BA925

SUÁREZ, J. L., GARCÍA, S. and HERRERA, F. (2021). A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. *Neurocomputing* **425** 300–322.

SUN, Y. and GENTON, M. G. (2011). Functional boxplots. *J. Comput. Graph. Statist.* **20** 316–334. MR2847798 https://doi.org/10.1198/jcgs.2011.09224

TEIXEIRA, L. V., ASSUNÇÃO, R. M. and LOSCHI, R. H. (2019). Bayesian space-time partitioning by sampling and pruning spanning trees. *J. Mach. Learn. Res.* **20** Paper No. 85, 35 pp. MR3960939

VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27** 1413–1432. MR3647105 https://doi.org/10.1007/s11222-016-9696-4

VINER, R. M., RUSSELL, S. J., CROKER, H., PACKER, J., WARD, J., STANSFIELD, C., MYTTON, O., BONELL, C. and BOOY, R. (2020). School closure and management practices during coronavirus outbreaks including COVID-19: A rapid systematic review. *Lancet Child Adolesc. Health* **4** 397–404. https://doi.org/10.1016/S2352-4642(20)30095-X

WARREN, M. S. and SKILLMAN, S. W. (2020). Mobility changes in response to COVID-19. Preprint. Available at arXiv:2003.14228.

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

WEILL, J. A., STIGLER, M., DESCHENES, O. and SPRINGBORN, M. R. (2020). Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proc. Natl. Acad. Sci. USA* **117** 19658–19660.

XING, E. P., JORDAN, M. I., RUSSELL, S. J. and NG, A. Y. (2003). Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems* 521–528.

YEUNG, K. Y., FRALEY, C., MURUA, A., RAFTERY, A. E. and RUZZO, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17** 977–987.

ZHANG, B., SANG, H., LUO, Z. T. and HUANG, H. (2023). Supplement to "Bayesian clustering of spatial functional data with application to a human mobility study during COVID-19." https://doi.org/10.1214/22-AOAS1643SUPPA, https://doi.org/10.1214/22-AOAS1643SUPPB

ZHANG, Z., LIM, C. Y., MAITI, T. and KATO, S. (2016). Spatial clustering of curves with functional covariates: A Bayesian partitioning model with application to spectra radiance in climate study. Preprint. Available at arXiv:1604.00059.

ZHOU, Q. and GUAN, Y. (2019). Fast model-fitting of Bayesian variable selection regression using the iterative complex factorization algorithm. *Bayesian Anal.* **14** 573–594. MR3934098 https://doi.org/10.1214/18-BA1120

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 606–620
https://doi.org/10.1214/22-AOAS1644

DESIGNING EXPERIMENTS FOR ESTIMATING AN APPROPRIATE OUTLET SIZE FOR A SILO TYPE PROBLEM

BY JESUS LOPEZ-FIDALGO[1,a], CATERINA MAY[2,b] AND JOSE ANTONIO MOLER[3,c]

[1]*Instituto de Ciencia de los Datos e Inteligencia Artificial, Universidad de Navarra,* [a]*fidalgo@unav.es*
[2]*Dipartimento DiSEI, Università degli Studi del Piemonte Orientale,* [b]*caterina.may@uniupo.it*
[3]*Departamento de Estadística, Informática y Matemáticas, Universidad Pública de Navarra,* [c]*jmoler@unavarra.es*

Jam formation is a problem that may occur when granular material is discharged by gravity from a silo. The estimation of the minimum outlet size, which guarantees that the time to the next jamming event is long enough, can be crucial in the industry. The time is modeled by an exponential distribution with two unknown parameters, and this goal translates to precise estimation of a nonlinear transformation of the parameters. We obtain c-optimum experimental designs with that purpose, applying the graphic Elfving method. Because the optimal experimental designs depend on the nominal values of the parameters, we conduct a sensitivity analysis on our dataset. Finally, a simulation study checks the performance of the approximations, first with the Fisher Information matrix, then with the linearization of the function to be estimated. The results are useful for experimenting in a laboratory and then translating the results to a real scenario. From the application we develop a general methodology for estimating a one-dimensional transformation of the parameters of a nonlinear model.

REFERENCES

ABLE, R., OTHEN, S. and NEDDERMAN, R. (1996). The exit time distribution during the batch discharge of a cylindrical bunker. *Chem. Eng. Sci.* **51** 4605–4610.

AMO-SALAS, M., DELGADO-MÁRQUEZ, E. and LÓPEZ-FIDALGO, J. (2016). Optimal experimental designs in the flow rate of particles. *Technometrics* **58** 269–276. MR3488305 https://doi.org/10.1080/00401706.2015.1042169

AMO-SALAS, M., DELGADO-MÁRQUEZ, E., FILOVÁ, L. and LÓPEZ-FIDALGO, J. (2016). Optimal designs for model discrimination and fitting for the flow of particles. *Statist. Papers* **57** 875–891. MR3571178 https://doi.org/10.1007/s00362-016-0792-5

ARTONI, R., SANTOMASO, A. and CANU, P. (2009). Simulation of dense granular flows: Dynamics of wall stress silos. *Chem. Eng. Sci.* **64** 4040–4050.

ATKINSON, A. C., DONEV, A. N. and TOBIAS, R. D. (2007). *Optimum Experimental Designs, with SAS. Oxford Statistical Science Series* **34**. Oxford Univ. Press, Oxford. MR2323647

BELL, T. (2005). Challenges in the scale-up of particulate processes—an industrial perspective. *Powder Technol.* **150** 60–71.

BROWN, M. (1990). Error bounds for exponential approximations of geometric convolutions. *Ann. Probab.* **18** 1388–1402. MR1062073

CARSON, J. and CRAIG, D. (2015). Silo design codes: Their limits and inconsistencies. *Proc. Eng.* **102** 647–656.

CARSON, J. W. (2008). *Bulk Solids Handling. Equipment Selection and Operation* (D. McGlinchey, ed.) Blackwell Publishing, Oxford.

CHEN, W. and ROBERTS, A. W. (2018). A modified flowability classification model for moist and cohesive bulk solids. *Powder Technol.* **325** 639–650.

ELFVING, G. (1952). Optimum allocation in linear regression theory. *Ann. Math. Stat.* **84** 44002–1–44002–6.

FEDOROV, V. V. (1972). *Theory of Optimal Experiments*. Academic Press, New York. MR0403103

FITZPATRICK, J. J., BARRINGER, S. A. and IQBAL, T. (2004). Flow property measurement of food powders and sensitivity of Jenike's Hopper design methodology to the measured values. *J. Food Eng.* **61** 399–405.

Key words and phrases. Bulk solid storage, jam formation, nonlinear heteroscedastic model, optimal design of experiments.

GELLA, D., ZURIGUEL, I. and MAZA, D. (2018). Decoupling geometrical and kinematic contributions to the silo clogging process. *Phys. Rev. Lett.* **121** 138001.

GOLSHAM, S., ESGANDARI, B., ZARGHAMI, R., BLAIS, B. and SALEH, K. (2020). Experimental and DEM studies of velocitiy profiles and residence time distribution of non-spherical particles in silos. *Powder Technol.* **373** 510–521.

JANDA, A., ZURIGUEL, I., GARCIMARTIN, A., PUGNALONI, L. A. and MAZA, D. (2008). Jamming and critical outlet size in the discharge of a two-dimensional silo. *Europhys. Lett.* **23** 255–262.

JENIKE, A. W. (1961). *Gravity Flow of Bulk Solids*. Bulletin of the Univ. Utah.

LI, J., LANGSTON, P. A., WEBB, C. and DYAKOWSKI, T. (2004). Flow of sphero-disc particles in rectangular hoppers-a DEM and experimental comparison in 3D). *Chem. Eng. Sci.* **59** 5917–5929.

LOPEZ-FIDALGO, J., MAY, C. and MOLER, J. A (2023). Supplement to "Designing experiments for estimating an appropriate outlet size for a silo type problem." https://doi.org/10.1214/22-AOAS1644SUPP

LÓPEZ-FIDALGO, J. and RODRÍGUEZ-DÍAZ, J. M. (2004). Elfving's method for m-dimensional models. *Metrika* **59** 235–244. MR2066727 https://doi.org/10.1007/s001840300281

MICCIO, F., BARLETTA, D. and POLETTO, M. (2013). Flow properties and arching behavior of biomass particulate solids. *Powder Technol.* **235** 312–321.

MITRA, H., PUSHPADASS, H. A., ELJEEVA, M., FRANKLIN, E., KINGSBY, R. P., GHOROI, C., NATH, S. and BATTULA, S. (2017). Influence of moisture content on the flow properties of basundi mix. *Powder Technol.* **312** 133–143.

NEDDERMAN, R. M. (1992). *Statics and Kinematics of Granular Materials*. Cambridge Univ. Press, Cambridge.

RIGBY, R. A., STASINOPOULOS, M. D., HELLER, G. Z. and DE BASTIANI, F. (2019). *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*, 1st ed. CRC Press, Boca Raton, FL.

RIVAS-LÓPEZ, M. J., LÓPEZ-FIDALGO, J. and DEL CAMPO, R. (2014). Optimal experimental design for accelerated failure time with Type I and random censoring. *Biom. J.* **56** 819–837. MR3258098 https://doi.org/10.1002/bimj.201300209

ROTTER, J. M. (2001). *Guide for the Economic Design of Circular Metal Silos*. Taylor & Francis Routledge, London.

SALEH, K., GOLSHAN, S. and ZARGHAMI, R. (2018). A review on gravity floe of free-flowing granular solids in silos—basics and practical aspects. *Chem. Eng. Sci.* **192** 1011–1035.

SAMUELSSON, R., LARSSON, S. H., THYREL, M. and LESTANDER, T. A. (2012). Moisture content and storage time influence the binding mechanisms in biofuel wood pellets. *Appl. Energy* **99** 109–115.

SCHULZE, D. (2008). *Powders and Bulk Solids*. Springer, Berlin.

TO, K. (2005). Jamming transition in two-dimensional hoppers and silos. *Phys. Rev. E* **71** 060301.

ZURIGUEL, I., GARCIMARTÍN, A., MAZA, D., PUGNALONI, L. A. and PASTOR, J. M. (2005). Jamming during the discharge of granular matter from a silo. *Phys. Rev. E* **71** 051303.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 621–640
https://doi.org/10.1214/22-AOAS1645

TEAM: A MULTIPLE TESTING ALGORITHM ON THE AGGREGATION TREE FOR FLOW CYTOMETRY ANALYSIS

BY JOHN A. PURA[1,a], XUECHAN LI[2,b], CLIBURN CHAN[2,c] AND JICHUN XIE[2,d]

[1]*Center of Innovation to Accelerate Discovery and Practice Transformation, Durham Veterans Affairs Medical Center,* [a]*john.pura@duke.edu*

[2]*Department of Biostatistics and Bioinformatics, Duke University School of Medicine,* [b]*xuechan.li@duke.edu,* [c]*cliburn.chan@duke.edu,* [d]*jichun.xie@duke.edu*

In immunology studies, flow cytometry is a commonly used multivariate single-cell assay. One key goal in flow cytometry analysis is to detect the immune cells responsive to certain stimuli. Statistically, this problem can be translated into comparing two protein expression probability density functions (PDFs) before and after the stimulus; the goal is to pinpoint the regions where these two PDFs differ. Further screening of these differential regions can be performed to identify enriched sets of responsive cells. In this paper we model identifying differential density regions as a multiple testing problem. First, we partition the sample space into small bins. In each bin we form a hypothesis to test the existence of differential PDFs. Second, we develop a novel multiple testing method, called TEAM (testing on the aggregation tree method), to identify those bins that harbor differential PDFs while controlling the false discovery rate (FDR) under the desired level. TEAM embeds the testing procedure into an aggregation tree to test from fine- to coarse-resolution. The procedure achieves the statistical goal of pinpointing density differences to the smallest possible regions. TEAM is computationally efficient, capable of analyzing large flow cytometry data sets in much shorter time compared with competing methods. We applied TEAM and competing methods on a flow cytometry data set to identify T cells responsive to the cytomegalovirus (CMV)-pp65 antigen stimulation. With additional downstream screening, TEAM successfully identified enriched sets containing monofunctional, bifunctional, and polyfunctional T cells. Competing methods either did not finish in a reasonable time frame or provided less interpretable results. Numerical simulations and theoretical justifications demonstrate that TEAM has asymptotically valid, powerful, and robust performance. Overall, TEAM is a computationally efficient and statistically powerful algorithm that can yield meaningful biological insights in flow cytometry studies.

REFERENCES

ANTONIADIS, A., GLAD, I. K. and MOHAMMED, H. (2015). Local comparison of empirical distributions via nonparametric regression. *J. Stat. Comput. Simul.* **85** 2384–2405. MR3346029 https://doi.org/10.1080/00949655.2014.929133

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

BOGOMOLOV, M., PETERSON, C. B., BENJAMINI, Y. and SABATTI, C. (2021). Hypotheses on a tree: New error rates and testing strategies. *Biometrika* **108** 575–590. MR4298765 https://doi.org/10.1093/biomet/asaa086

DMITRIENKO, A. and TAMHANE, A. C. (2011). Mixtures of multiple testing procedures for gatekeeping applications in clinical trials. *Stat. Med.* **30** 1473–1488. MR2828882 https://doi.org/10.1002/sim.4008

DMITRIENKO, A. and TAMHANE, A. C. (2013). General theory of mixture procedures for gatekeeping. *Biom. J.* **55** 402–419. MR3049194 https://doi.org/10.1002/bimj.201100258

DUONG, T. (2013). Local significant differences from nonparametric two-sample tests. *J. Nonparametr. Stat.* **25** 635–645. MR3174288 https://doi.org/10.1080/10485252.2013.810217

Key words and phrases. Flow cytometry, multiple testing, aggregation tree, distribution difference, false discovery proportion (FDP).

GOEMAN, J. J. and FINOS, L. (2012). The inheritance procedure: Multiple testing of tree-structured hypotheses. *Stat. Appl. Genet. Mol. Biol.* **11** Art. 11, 20. MR2924200 https://doi.org/10.1515/1544-6115.1554

GUO, W., LYNCH, G. and ROMANO, J. P. (2018). A new approach for large scale multiple testing with application to FDR control for graphically structured hypotheses. Preprint. Available at arXiv:1812.00258.

HAHNE, F., KHODABAKHSHI, A. H., BASHASHATI, A., WONG, C.-J., GASCOYNE, R. D., WENG, A. P., SEYFERT-MARGOLIS, V., BOURCIER, K., ASARE, A. et al. (2010). Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A* **77** 121–131. https://doi.org/10.1002/cyto.a.20823

LI, Y., HU, Y.-J. and SATTEN, G. A. (2022). A bottom-up approach to testing hypotheses that have a branching tree dependence structure, with error rate control. *J. Amer. Statist. Assoc.* **117** 664–677. MR4436304 https://doi.org/10.1080/01621459.2020.1799811

LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* **41** 2948–2978. MR3161453 https://doi.org/10.1214/13-AOS1169

LIU, W. and SHAO, Q.-M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *Ann. Statist.* **42** 2003–2025. MR3262475 https://doi.org/10.1214/14-AOS1249

MEIJER, R. J. and GOEMAN, J. J. (2015). A multiple testing method for hypotheses structured in a directed acyclic graph. *Biom. J.* **57** 123–143. MR3298222 https://doi.org/10.1002/bimj.201300253

PURA, J. A., LI, X., CHAN, C. AND XIE, J. (2023). SUPPLEMENT TO "TEAM: A MULTIPLE TESTING ALGORITHM ON THE AGGREGATION TREE FOR FLOW CYTOMETRY ANALYSIS." HTTPS://DOI.ORG/10.1214/22-AOAS1645SUPP

RAMDAS, A., CHEN, J., WAINWRIGHT, M. J. and JORDAN, M. I. (2019). A sequential algorithm for false discovery rate control on directed acyclic graphs. *Biometrika* **106** 69–86. MR3912384 https://doi.org/10.1093/biomet/asy066

ROEDERER, M. and HARDY, R. R. (2001). Frequency difference gating: A multivariate method for identifying subsets that differ between samples. *Cytometry, Part A* **45** 56–64.

ROEDERER, M., MOORE, W., TREISTER, A., HARDY, R. R. and HERZENBERG, L. A. (2001). Probability binning comparison: A metric for quantitating multivariate distribution differences. *Cytometry, Part A* **45** 47–55.

SEDER, R. A., DARRAH, P. A. and ROEDERER, M. (2008). T-cell quality in memory and protection: Implications for vaccine design. *Nat. Rev., Immunol.* **8** 247–58. https://doi.org/10.1038/nri2274

SORIANO, J. and MA, L. (2017). Probabilistic multi-resolution scanning for two-sample differences. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 547–572. MR3611759 https://doi.org/10.1111/rssb.12180

STAATS, J. S., ENZOR, J. H., SANCHEZ, A. M., ROUNTREE, W., CHAN, C., JAIMES, M., CHAN, R. C.-F., GAUR, A., DENNY, T. N. et al. (2014). Toward development of a comprehensive external quality assurance program for polyfunctional intracellular cytokine staining assays. *J. Immunol. Methods* **409** 44–53.

R CORE TEAM (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

VALITUTTI, S., MÜLLER, S., SALIO, M. and LANZAVECCHIA, A. (1997). Degradation of T cell receptor (TCR)-CD3-zeta complexes after antigenic stimulation. *J. Exp. Med.* **185** 1859–1864. https://doi.org/10.1084/jem.185.10.1859

WALTHER, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.* **38** 1010–1033. MR2604703 https://doi.org/10.1214/09-AOS732

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. https://doi.org/10.1186/s12859-019-2599-6

XIE, J. and LI, R. (2018). False discovery rate control for high dimensional networks of quantile associations conditioning on covariates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 1015–1034. MR3874308 https://doi.org/10.1111/rssb.12288

YEKUTIELI, D. (2008). Hierarchical false discovery rate-controlling methodology. *J. Amer. Statist. Assoc.* **103** 309–316. MR2420235 https://doi.org/10.1198/016214507000001373

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 641–679
https://doi.org/10.1214/22-AOAS1646

ALPHA: AUDIT THAT LEARNS FROM PREVIOUSLY HAND-AUDITED BALLOTS

BY PHILIP B. STARK[a]

Department of Statistics, University of California, Berkeley, [a]*stark@stat.berkeley.edu*

A risk-limiting election audit (RLA) offers a statistical guarantee: if the reported electoral outcome is incorrect, the audit has a known maximum chance (the risk limit) of not correcting it before it becomes final. BRAVO (Lindeman, Stark and Yates (In *Proceedings of the 2011 Electronic Voting Technology Workshop/Workshop on Trustworthy Elections (EVT/WOTE'11)* (2012) USENIX)), based on Wald's sequential probability ratio test for the Bernoulli parameter, is the simplest and most widely tried method for RLAs, but it has limitations. It cannot accommodate sampling without replacement or stratified sampling which can improve efficiency and are sometimes required by law. It applies only to ballot-polling audits which are less efficient than comparison audits. It applies to plurality, majority, supermajority, proportional representation, and instant-runoff voting (IRV, using RAIRE (Blom, Stuckey and Teague (In *Electronic Voting* (2018) 17–34 Springer))) but not to other social choice functions for which there are RLA methods. And while BRAVO has the smallest expected sample size among sequentially valid ballot-polling-with-replacement methods when the reported vote shares are exactly correct, it can require arbitrarily large samples when the reported reported winner(s) really won but the reported vote shares are incorrect. ALPHA is a simple generalization of BRAVO that: (i) works for sampling with and without replacement, with and without weights, with and without stratification, and for Bernoulli sampling; (ii) works not only for ballot polling but also for ballot-level comparison, batch polling, and batch-level comparison audits; (iii) works for all social choice functions covered by SHANGRLA (Stark (In *Financial Cryptography and Data Security* (2020) Springer)), including approval voting, STAR-Voting, proportional representation schemes, such as D'Hondt and Hamilton, IRV, Borda count, and all scoring rules, and (iv) in situations where both ALPHA and BRAVO apply, requires smaller samples than BRAVO when the reported vote shares are wrong but the outcome is correct—five orders of magnitude in some examples. ALPHA includes the family of betting martingale tests in RiLACS (Waudby-Smith, Stark and Ramdas (In *Electronic Voting. E-Vote-ID 2021* (2021) Springer)) with a different betting strategy parametrized as an estimator of the population mean and explicit flexibility to accommodate sampling weights and population bounds that change with each draw. A Python implementation is provided.

REFERENCES

APPEL, A. W., DEMILLO, R. and STARK, P. B. (2020). Ballot-marking devices cannot assure the will of the voters. *Election Law Journal*: *Rules*, *Politics*, *and Policy* **19**. https://doi.org/10.1089/elj.2019.0619

APPEL, A. W. and STARK, P. B. (2020). Evidence-based elections: Create a meaningful paper trail, then audit. *Georgetown Law Technology Review* **4.2** 523–541. https://georgetownlawtechreview.org/wp-content/uploads/2020/07/4.2-p523-541-Appel-Stark.pdf.

AMERICAN STATISTICAL ASSOCIATION (2010). American Statistical Association statement on risk-limiting post-election audits. Technical Report, ASA. https://www.amstat.org/asa/files/pdfs/POL-ASARecommendsRisk-LimitingAudits.pdf, Retrieved 17 August 2022.

BAÑUELOS, J. H. and STARK, P. B. (2012). Limiting Risk by Turning Manifest Phantoms into Evil Zombies Technical Report arXiv.org. https://arxiv.org/abs/1207.3413, Retrieved 17 August 2022.

Key words and phrases. Risk-limiting audit, elections, supermartingale test, Ville's Inequality, SHANGRLA.

BLOM, M., STUCKEY, P. J. and TEAGUE, V. J. (2018). Ballot-polling risk limiting audits for IRV elections. In *Electronic Voting* (R. Krimmer, M. Volkamer, V. Cortier, R. Goré, M. Hapsara, U. Serdült and D. Duenas-Cid, eds.) 17–34. Springer, Cham.

GLAZER, A., SPERTUS, J. and STARK, P. B. (2021). More style, less work: Card-style data decrease risk-limiting audit sample sizes. *Digital Threats*: *Research and Practice* **2** 1–15. https://doi.org/10.1145/3457907

HIGGINS, M. J., RIVEST, R. L. and STARK, P. B. (2011). Sharper p-values for stratified post-election audits. *Statistics*, *Politics*, *and Policy* **2**.

HUANG, A., RIVEST, R. L., STARK, P. B., TEAGUE, V. and VUKCEVIC, D. (2020). A unified evaluation of two-candidate ballot-polling election auditing methods. In *Proceedings of E-Vote ID* 2020. *Lecture Notes in Computer Science* (R. Krimmer, M. Volkamer, B. Beckert, R. Küsters, O. Kulyk, D. Duenas-Cid and M. Solvak, eds.) Springer, Cham.

KELLY, J. L. JR. (1956). A new interpretation of information rate. *Bell Syst*. *Tech*. *J*. **35** 917–926. MR0090494 https://doi.org/10.1002/j.1538-7305.1956.tb03809.x

LINDEMAN, M., STARK, P. B. and YATES, V. (2012). BRAVO: Ballot-polling risk-limiting audits to verify outcomes. In *Proceedings of the* 2011 *Electronic Voting Technology Workshop/Workshop on Trustworthy Elections* (*EVT/WOTE*'11) USENIX, Bellevue, WA.

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE (2018). *Securing the Vote*: *Protecting American Democracy*. The National Academies Press, Washington, DC. https://doi.org/10.17226/25120

OTTOBONI, K., BERNHARD, M., HALDERMAN, J. A., RIVEST, R. L. and STARK, P. B. (2020). Bernoulli ballot polling: A manifest improvement for risk-limiting audits. In *Financial Cryptography and Data Security*. *Lecture Notes in Computer Science* **11599** 226–241. Springer, Cham. MR4359233 https://doi.org/10.1007/978-3-030-43725-1_16

OTTOBONI, K., STARK, P. B., LINDEMAN, M. and MCBURNETT, N. (2018). Risk-limiting audits by stratified union-intersection tests of elections (SUITE). In *Electronic Voting*. *E-Vote-ID* 2018. *Lecture Notes in Computer Science* Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-00419-4_12.

RIVEST, R. L. (2017). ClipAudit: A Simple Risk-Limiting Post-Election Audit. https://arxiv.org/abs/1701.08312.

RIVEST, R. L. (2018). Bayesian Tabulation Audits: Explained and Extended. https://arxiv.org/abs/1801.00528.

RIVEST, R. L. and SHEN, E. (2012). A Bayesian method for auditing elections. In 2012 *Electronic Voting Technology/Workshop on Trustworthy Elections* (*EVT/WOTE*'12).

SPERTUS, J. and STARK, P. B. (2022). Sweeter than SUITE: Supermartingale stratified union-intersection tests of elections. In *Electronic Voting*, *E-VOTE-ID* 2022. *Lecture Notes in Computer Science* Springer, Cham.

STARK, P. B. (2008). Conservative statistical post-election audits. *Ann*. *Appl*. *Stat*. **2** 550–581. MR2524346 https://doi.org/10.1214/08-AOAS161

STARK, P. B. (2009a). Risk-limiting postelection audits: Conservative P-values from common probability inequalities. *IEEE Trans*. *Inform*. *Forensics Secur*. **4** 1005–1014. https://doi.org/10.1109/TIFS.2009.2034190

STARK, P. B. (2009b). CAST: Canvass audits by sampling and testing. *IEEE Trans*. *Inform*. *Forensics Secur*. **4** 708–717.

STARK, P. B. (2020). Sets of half-average nulls generate risk-limiting audits: SHANGRLA. In *Financial Cryptography and Data Security*. *Lecture Notes in Computer Science* **12063** (M. Bernhard, A. Bracciali, L. J. Camp, S. Matsuo, A. Maurushat, P. B. Rønne and M. Sala, eds.) Springer, Cham.

STARK, P. B. (2022). Non(c)esuch ballot-level risk-limiting audits for precinct-count voting systems. https://arxiv.org/abs/2207.01362.

STARK, P. B. (2023). Supplement to "ALPHA: Audit that learns from previously hand-audited ballots." https://doi.org/10.1214/22-AOAS1646SUPP

STARK, P. B. and EVANS, S. N. Inference about population means from sequential samples using martingales. https://github.com/pbstark/MartInf.

STARK, P. B. and TEAGUE, V. (2014). Verifiable European elections: Risk-limiting audits for D'Hondt and its relatives. *JETS*: *USENIX Journal of Election Technology and Systems* **3.1**.

STARK, P. B. and WAGNER, D. A. (2012). Evidence-based elections. *IEEE Secur*. *Priv*. **10** 33–41.

VILLE, J. (1939). *Etude Critique de la Notion de Collectif*. *Monographies des Probabilités* **3**. Gauthier-Villars, Paris.

VORA, P. L. (2019). Risk-Limiting Bayesian Polling Audits for Two Candidate Elections. http://arxiv.org/abs/1902.00999.

WALD, A. (1945). Sequential tests of statistical hypotheses. *Ann*. *Math*. *Stat*. **16** 117–186. MR0013275 https://doi.org/10.1214/aoms/1177731118

WAUDBY-SMITH, I. and RAMDAS, A. (2021). Estimating means of bounded random variables by betting. https://arxiv.org/abs/2010.09686.

WAUDBY-SMITH, I., STARK, P. B. and RAMDAS, A. (2021). RiLACS: Risk-limiting audits via confidence sequences. In *Electronic Voting*. *E-Vote-ID* 2021. *Lecture Notes in Computer Science* **12900** (R. Krimmer,

M. Volkamer, D. Duenas-Cid, O. Kulyk, P. Rønne, M. Solvak and M. Germann, eds.) Springer, Cham. https://doi.org/10.1007/978-3-030-86942-7\protect\T1\textunderscore9

ZAGÓRSKI, F., MCCLEARN, G., MORIN, S., MCBURNETT, N. and VORA, P. L. (2021). Minerva—an efficient risk-limiting ballot polling audit. In 30*th USENIX Security Symposium*. USENIX, Berkeley, CA.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 680–710
https://doi.org/10.1214/22-AOAS1647

MODEL SELECTION UNCERTAINTY AND STABILITY IN BETA REGRESSION MODELS: A STUDY OF BOOTSTRAP-BASED MODEL AVERAGING WITH AN EMPIRICAL APPLICATION TO CLICKSTREAM DATA

BY CORBAN ALLENBRAND[a] AND BEN SHERWOOD[b]

School of Business, University of Kansas, [a]*callenbrand@ku.edu,* [b]*ben.sherwood@ku.edu*

Statistical model development is a central feature of many scientific investigations with a vast methodological landscape. However, uncertainty in the model development process has received less attention and is frequently resolved nonrigorously through beliefs about generalizability, practical usefulness, and computational ease. This is particularly problematic in settings of abundant data, such as clickstream data, as model selection routinely admits multiple models and imposes a source of uncertainty, unacknowledged and unknown by many, on all postselection conclusions. Regression models, based on the beta distribution, are a class of nonlinear models, attractive because of their great flexibility and potential explanatory power, but have not been investigated from the standpoint of multimodel uncertainty and model averaging. For this reason a formalized tool that can combine model selection uncertainty and beta regression modeling is presented in this work. The tool combines bootstrap model averaging, model selection, and asymptotic theory to yield a procedure that can perform joint modeling of the mean and precision parameters, capture sources of variability in the data, and achieve more accurate claims of estimate precision, variable importance, generalization performance, and model stability. Practical utility of the tool is demonstrated through a study of model selection consistency and variable importance in average exit and bounce rate statistical models. This work emphasizes the necessity of a departure from the all-too-common practice of ignoring model selection uncertainty and introduces an accessible technique to handle frequently neglected aspects of the modeling pipeline.

REFERENCES

ALLENBRAND, C. and SHERWOOD, B. (2023). Supplement to "Model selection uncertainty and stability in beta regression models: A study of bootstrap-based model averaging with an empirical application to clickstream data." https://doi.org/10.1214/22-AOAS1647SUPP

ANDERSON, D. R., BURNHAM, K. P. and WHITE, G. C. (1994). AIC model selection in overdispersed capture–recapture data. *Ecology* **75** 1780–1793.

ANSARI, S. et al. (2002). Integrating E-commerce and data mining: Architecture and challenges. In *IEEE International Conference on Data Mining*.

AUGUSTIN, N., SAUERBREI, W. and SCHUMACHER, M. (2005). The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Stat. Model.* **5** 95–118. MR2140338 https://doi.org/10.1191/1471082X05st089oa

BAUMEISTER, R. F. (2002). Yielding to temptation: Self-control failure, impulsive purchasing, and consumer behavior. *J. Consum. Res.* **28** 670–676.

BIAU, G. and SCORNET, E. (2016). A random forest guided tour. *TEST* **25** 197–227. MR3493512 https://doi.org/10.1007/s11749-016-0481-7

BOOTH, D. L. and JANSEN, B. J. (2008). A review of methodologies for analyzing websites. In *Handbook of Research on Web Log Analysis* (B. J. Jansen, A. Spink and I. Taska, eds.) 143–164. IGI, Hershey, PA.

BREIMAN, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Amer. Statist. Assoc.* **87** 738–754. MR1185196

Key words and phrases. Beta distribution, generalized linear model, variable selection, model combination, corrected AIC, rates and proportions.

BREIMAN, L. and FREEDMAN, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78** 131–136. MR0696857

BUCHHOLZ, A., HOLLÄNDER, N. and SAUERBREI, W. (2008). On properties of predictors derived with a two-step bootstrap model averaging approach—A simulation study in the linear regression model. *Comput. Statist. Data Anal.* **52** 2778–2793. MR2419541 https://doi.org/10.1016/j.csda.2007.10.007

BUCKLAND, S. T., BURNHAM, K. P. and AUGUSTIN, N. H. (1997). Model selection: An integral part of inference. *Biometrics* **53** 603–618.

BURNHAM, K. P. and ANDERSON, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer, New York. MR1919620

CEPEDA-CUERVO, E. (2015). Beta regression models: Joint mean and variance modeling. *J. Stat. Theory Pract.* **9** 134–145. MR3275536 https://doi.org/10.1080/15598608.2014.890983

CEPEDA-CUERVO, E. and GAMERMAN, D. (2005). Bayesian methodology for modeling parameters in the two parameter exponential family. *Estadística* **57** 93–105. MR2406467

CHAFFEY, D. and PATRON, M. (2012). From web analytics to digital marketing optimization: Increasing the commerical value of digital analytics. *J. Direct Data Digit. Mark. Pract.* **14** 30–45.

CHALMERS, R. P., PEK, J. and LIU, Y. (2017). Profile-likelihood confidence intervals in item response theory models. *Multivar. Behav. Res.* **52** 533–550. https://doi.org/10.1080/00273171.2017.1329082

CHATERJEE, P., HOFFMAN, D. L. and NOVAK, T. P. (2003). Modeling the clickstream: Implications for web-based advertising efforts. *Mark. Sci.* **22** 437–454.

CHATFIELD, C. (1995). Model uncertainty, data mining and statistical inference. *J. Roy. Statist. Soc. Ser. A* **158** 419–466.

CHO, C. H., KANG, J. and CHEON, H. J. (2006). Online shopper hesitation. *Cyberpsychol. Behav.* **9** 261–274.

CIALDINI, R. B. (2001). Harnessing the science of persuasion. *Harv. Bus. Rev.* **10** 72–79.

CLAESKENS, G. and HJORT, N. L. (2008). *Model Selection and Model Averaging. Cambridge Series in Statistical and Probabilistic Mathematics* **27**. Cambridge Univ. Press, Cambridge. MR2431297 https://doi.org/10.1017/CBO9780511790485

COPAS, J. B. (1983). Regression, prediction and shrinkage. *J. Roy. Statist. Soc. Ser. B* **45** 311–354. MR0737642

CRIBARI-NETO, F. and ZEILEIS, A. (2010). Beta regression in R. *J. Stat. Softw.* **34** 1–24.

DE BIN, R., JANITZA, S., SAUERBREI, W. and BOULESTEIX, A.-L. (2016). Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometrics* **72** 272–280. MR3500596 https://doi.org/10.1111/biom.12381

DICICCIO, T. J. and EFRON, B. (1996). Bootstrap confidence intervals. *Statist. Sci.* **11** 189–228. MR1436647 https://doi.org/10.1214/ss/1032280214

EDITOR GOOGLE ANALYTICS (2020). Exit rate vs. bounce rate. Available at https://support.google.com/analytics/answer/2525491?hl=en. Accessed: Nov 20, 2020.

EMMERT-STREIB, F., YANG, Z., FENG, H., TRIPATHI, S. and DEHMER, M. (2020). An introductory review of deep learning for prediction models with big data. *Front. Artif. Intell.* **3** 4.

ESPINHEIRA, P. L., FERRARI, S. L. P. and CRIBARI-NETO, F. (2008a). On beta regression residuals. *J. Appl. Stat.* **35** 407–419. MR2420486 https://doi.org/10.1080/02664760701834931

ESPINHEIRA, P. L., FERRARI, S. L. P. and CRIBARI-NETO, F. (2008b). Influence diagnostics in beta regression. *Comput. Statist. Data Anal.* **52** 4417–4431. MR2432471 https://doi.org/10.1016/j.csda.2008.02.028

FAN, J., HAN, F. and LIU, H. (2014). Challenges of big data analysis. *Nat. Sci. Rev.* **1** 293–314.

FERRARI, S. L. P. and CRIBARI-NETO, F. (2004). Beta regression for modelling rates and proportions. *J. Appl. Stat.* **31** 799–815. MR2095753 https://doi.org/10.1080/0266476042000214501

GOLDFARB, A. and LU, Q. (2006). Household-specific regressions using clickstream data. *Statist. Sci.* **21** 247–255. MR2324084 https://doi.org/10.1214/088342306000000150

HANSEN, B. J. (2009). *Understanding User-Web Interactions via Web Analytics: Synthesis Lectures on Information Concepts*. Morgan and Claypool, Williston, VT.

HASAN, B. (2016). Perceived irritation in online shopping: The impact of website design characteristics. *Comput. Hum. Behav.* **54** 224–230.

HODGES, J. S. (1987). Uncertainty, policy analysis, and statistics. *Stat. Soc.* **2** 259–291.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417. MR1765176 https://doi.org/10.1214/ss/1009212519

HOWARD, J. A. and SHEITH, J. N. (1969). *The Theory of Buyer Behavior*. Wiley, New York.

HURVICH, C. M. and TSAI, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76** 297–307. MR1016020 https://doi.org/10.1093/biomet/76.2.297

KIESCHNICK, R. and MCCULLOUGH, B. D. (2003). Regression analysis of variates observed on $(0,1)$: Percentages, proportions and fractions. *Stat. Model.* **3** 193–213. MR2005473 https://doi.org/10.1191/1471082X03st053oa

KUHA, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociol. Methods Res.* **33** 188–229. MR2086348 https://doi.org/10.1177/0049124103262065

LEAMER, E. E. (1969). *Specification Searches*: *Ad Hoc Inference with Nonexperimental Data*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0471118

LEHMANN, E. L. (1983). *Theory of Point Estimation*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. MR0702834 https://doi.org/10.1007/978-1-4757-2769-2

LIU, F. and KONG, Y. (2015a). zombi: An R package for Bayesian inference for beta regression and zero/one inflated beta regression. *R J.* **7** 34–51.

LIU, F. and KONG, Y. (2015b). Bayesian inference for zero/one inflated beta regression model. R package version 1.3.3.

LIU, F. and LI, Q. (2016). A Bayesian model for joint analysis of multivariate repeated measures and time to event data in crossover trials. *Stat. Methods Med. Res.* **25** 2180–2192. MR3553333 https://doi.org/10.1177/0962280213519594

MARIN, M., ROJAS, J. and JAIMES, D. (2014). Bayesian beta regression: Joint mean and precision modeling. R package version 1.2.

MEHRZADI, D. and FEITELSON, D. G. (2012). On extracting session data from activity logs. In *Proceedings of the* 5*th Annual International Systems and Storage Conference*.

MILLER, A. J. (1990). *Subset Selection in Regression*. *Monographs on Statistics and Applied Probability* **40**. CRC Press, London. MR1072361 https://doi.org/10.1007/978-1-4899-2939-6

MOE, W. W. (2006). An empirical two-stage choice model with varying decision rules applied to Internet clickstream data. *Mark. Res.* **43** 680–692.

MOORE, W. W. and FADER, P. S. (2004). Capturing visit behavior in clickstream data. *J. Interact. Mark.* **18** 5–19.

MORAL-BENITO, E. (2015). Model averaging in economics: An overview. *J. Econ. Surv.* **29** 46–75.

MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. MR1803168 https://doi.org/10.2307/2669386

PAI, D., SHARANG, A., YADAGIRI, M. M. and AGRAWAL, S. (2014). Modelling similarity using click-stream data: A supervised approach. In *Web Information System Engineering—WISE*2014. *Lecture Notes in Computer Science* **8786**. Springer, Cham.

PAOLINO, P. (2001). Maximum likelihood estimation of models with beta-distribted dependent variables. *Polit. Anal.* **9** 325–346.

PARK, C. H. and KIM, Y. G. (2010). Identifying key factors affecting consumer purchase behavior in an online shopping context. *Int. J. Retail Distrib. Manag.* **3112** 16–29.

RAMANATHAN, V. et al. (2018). E-commerce trends—A service enterprise engineering perspective. Technical report, Penn State College of Engineering.

ROCHA, A. V. and CRIBARI-NETO, F. (2009). Beta autoregressive moving average models. *TEST* **18** 529–545. MR2566415 https://doi.org/10.1007/s11749-008-0112-z

ROCHA, A. V. and SIMAS, A. B. (2011). Influence diagnostics in a general class of beta regression models. *TEST* **20** 95–119. MR2806312 https://doi.org/10.1007/s11749-010-0189-z

SAKAR, C. O., POLAT, S. O., KATIRCIOGLU, M. and KASTRO, Y. (2019). Real-time prediction of online shopper's purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Comput. Appl.* **31** 6893–6908.

SENECAL, S., KALCZYNSKI, P. J. and NANTEL, J. (2005). Consumer's decision-making process and their online shopping behavior: A clickstream analysis. *J. Bus. Res.* **58** 1599–1608.

SIMAS, A. B., BARRETO-SOUZA, W. and ROCHA, A. V. (2010). Improved estimators for a general class of beta regression models. *Comput. Statist. Data Anal.* **54** 348–366. MR2756431 https://doi.org/10.1016/j.csda.2009.08.017

SMITHSON, M. and VERKUILEN, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol. Methods* **11** 54–71. https://doi.org/10.1037/1082-989X.11.1.54

SWINYARD, W. R. and SMITH, S. M. (2003). Why people (don't) shop online: A lifestyle study of the Internet consumer. *Psychol. Mark.* **20** 567–597.

THELWALL, M. (2010). *Introduction to Webometrics*. Morgan and Claypool, Williston, VT.

TOBIN, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26** 24–36. MR0090462 https://doi.org/10.2307/1907382

VOHS, K. D. and FABER, R. J. (2007). Spent resources: Self-regulatory resource availability affects impulse buying. *J. Consum. Res.* **33** 537–547.

WANG, G., ZHANG, X., TANG, S., ZHENG, H. and ZHAO, B. (2016). Unsupervised clustering for user behavior analysis. In *Proceedings of the* 2016 *CHI Conference on Human Factors in Computing Systems*.

WARTON, D. I. and HUI, F. K. C. (2011). The arcsine is asinine: The analysis of proportions in ecology. *Ecology* **92** 3–10.

YE, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93** 120–131. MR1614596 https://doi.org/10.2307/2669609

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 711–739
https://doi.org/10.1214/22-AOAS1648

DO FORECASTS OF BANKRUPTCY CAUSE BANKRUPTCY? A MACHINE LEARNING SENSITIVITY ANALYSIS

BY DEMETRIOS PAPAKOSTAS[1,a], P. RICHARD HAHN[1,b], JARED MURRAY[2,c], FRANK ZHOU[3,d] AND JOSEPH GERAKOS[4,e]

[1]*School of Mathematical and Statistical Sciences, Arizona State University,* [a]*dpapakos@asu.edu,* [b]*prhahn@asu.edu*
[2]*Department of Information, Risk and Operations Management, The University of Texas at Austin,* [c]*jared.murray@mccombs.utexas.edu*
[3]*Accounting Department, The Wharton School, University of Pennsylvania,* [d]*szho@wharton.upenn.edu*
[4]*Tuck School of Business, Dartmouth College,* [e]*Joseph.J.Gerakos@tuck.dartmouth.edu*

It is widely speculated that auditors' public forecasts of bankruptcy are, at least in part, self-fulfilling prophecies in the sense that they actually cause bankruptcies that would not have otherwise occurred. This conjecture is hard to prove, however, because the strong association between bankruptcies and bankruptcy forecasts could simply indicate that auditors are skillful forecasters with unique access to highly predictive covariates. In this paper we investigate the causal effect of bankruptcy forecasts on bankruptcy using nonparametric sensitivity analysis. We contrast our analysis with two alternative approaches: a linear bivariate probit model with an endogenous regressor and a recently developed bound on risk ratios called E-values. Additionally, our machine learning approach incorporates a monotonicity constraint corresponding to the assumption that bankruptcy forecasts do not make bankruptcies less likely. Finally, a tree-based posterior summary of the treatment effect estimates allows us to explore which observable firm characteristics moderate the inducement effect.

REFERENCES

ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *J. Amer. Statist. Assoc.* **105** 493–505. MR2759929 https://doi.org/10.1198/jasa.2009.ap08746

ABADIE, A. and GARDEAZABAL, J. (2003). The economic costs of conflict: A case study of the Basque country. *Am. Econ. Rev.* **93** 113–132.

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

ALTONJI, J., ELDER, T. and TABER, C. (2005). An evaluation of instrumental variable strategies for estimating the effects of catholic schooling. *J. Hum. Resour.* **60** 791–821.

BAO, Y., KE, B., LI, B., YU, Y. J. and ZHANG, J. (2020). Detecting accounting fraud in publicly traded us firms using a machine learning approach. *J. Acc. Res.* **58** 199–235.

BERTOMEU, J., CHEYNEL, E., FLOYD, E. and PAN, W. (2021). Using machine learning to detect misstatements. *Rev. Acc. Stud.* **26** 468–519.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BROWN, N. C., CROWLEY, R. M. and ELLIOTT, W. B. (2020). What are you saying? Using topic to detect financial misreporting. *J. Acc. Res.* **58** 237–291.

CAMPBELL, J., HILSCHER, J. and SZILAGYI, J. (2008). In search of distress risk. *J. Finance* **63** 2899–2939.

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** 772–793.

CHEN, P., HE, S., MA, Z. and STICE, D. (2016). The information role of audit opinions in debt contracting. *J. Account. Econ.* **61** 121–144.

CHEN, T. and GUESTRIN, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*.

Key words and phrases. BART, causal inference, heterogeneous treatment effects, self-fulfilling prophecy, sensitivity analysis.

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

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

CORNFIELD, J., HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A. M., SHIMKIN, M. B. and WYNDER, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* **22** 173–203.

DEFOND, M., RAGHUNANDAN, K. and SUBRAMANYAM, K. (2002). Do non-audit service fees impair auditor independence? Evidence from going concern audit opinions. *J. Acc. Res.* **40** 1247–1274.

EVANS, W. N. and SCHWAB, R. (1995). Finishing high school and starting college: Do catholic schools make a difference? *Q. J. Econ.* **110** 941–974.

FELDMANN, D. and READ, W. (2013). Going concern opinions for bankrupt companies—impact of credit rating. *Manag. Audit. J.* **28** 345–363.

FREEDMAN, D. A. and SEKHON, J. S. (2010). Endogeneity in probit response models. *Polit. Anal.* **18** 138–150.

GERAKOS, J., HAHN, P. R., KOVRIJNYKH, A. and ZHOU, F. (2016). Prediction versus inducement and the informational efficiency of going concern opinions. Chicago Booth (working paper).

GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Stat.* **4** 641–649.

HAHN, P. R. and CARVALHO, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *J. Amer. Statist. Assoc.* **110** 435–448. MR3338514 https://doi.org/10.1080/01621459.2014.993077

HAHN, P. R., HE, J. and LOPES, H. F. (2019). Efficient sampling for Gaussian linear regression with arbitrary priors. *J. Comput. Graph. Statist.* **28** 142–154. MR3939378 https://doi.org/10.1080/10618600.2018.1482762

HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal.* **15** 965–1056. MR4154846 https://doi.org/10.1214/19-BA1195

HECKMAN, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica* **46** 931–959. MR0483259 https://doi.org/10.2307/1909757

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

IMBENS, G. W. (2014). Instrumental variables: An econometrician's perspective. *Statist. Sci.* **29** 323–358. MR3264545 https://doi.org/10.1214/14-STS480

IMBENS, G. W. and LEMIEUX, T. (2008). Regression discontinuity designs: A guide to practice. *J. Econometrics* **142** 615–635. MR2416821 https://doi.org/10.1016/j.jeconom.2007.05.001

JINDAL, N. and MCALISTER, L. (2015). The impacts of advertising assets and r&d assets on reducing bankruptcy risk. *Mark. Sci.* **34** 555–572.

KAPLAN, S. E. and WILLIAMS, D. D. (2013). Do going concern audit reports protect auditors from litigation? A simultaneous equations approach. *Account. Rev.* **88** 199–232.

LARCKER, D. F. and RUSTICUS, T. O. (2010). On the use of instrumental variables in accounting research. *J. Account. Econ.* **49** 186–205.

MANSKI, C. F. (2007). *Identification for Prediction and Decision*. Harvard Univ. Press, Cambridge, MA.

MAURER, M. (2020). Going-concern opinions may be poised for a comeback. *Wall St. J.*.

MENG, C.-L. and SCHMIDT, P. (1985). On the cost of partial observability in the bivariate probit model. *Internat. Econom. Rev.* **26** 71–85. MR0802706 https://doi.org/10.2307/2526528

NELDER, J. A. and MEAD, R. (1965). A simplex method for function minimization. *Comput. J.* **7** 308–313. MR3363409 https://doi.org/10.1093/comjnl/7.4.308

PAPAKOSTAS, D., HAHN, P. R, MURRAY, J., ZHOU, F. and GERAKOS, J. (2023). Supplement to "Do forecasts of bankruptcy cause bankruptcy? A machine learning sensitivity analysis." https://doi.org/10.1214/22-AOAS1648SUPPA, https://doi.org/10.1214/22-AOAS1648SUPPB

PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press, Cambridge. MR1744773

PENG, D. and VANDERWEELE, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27** 368–377. https://doi.org/10.1097/EDE.0000000000000457

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and non randomized studies. *J. Educ. Psychol.* **66** 688–701.

THISTLETHWAITE, D. L. and CAMPBELL, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *J. Educ. Psychol.* **51** 309.

VANDERWEELE, T. J. and DING, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Ann. Intern. Med.* **168** 268–274.

WOODY, C., CARVALHO, S., HAHN, P. R. and MURRAY, J. (2020). Estimating heterogeneous effects of continuous exposures using Bayesian tree ensembles: Revisiting the impact of abortion rates on crime. Preprint. Available at arXiv:2007.09845.

WOOLRIDGE, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. Massachusetts Institute of Technology, Cambridge, MA.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 740–760
https://doi.org/10.1214/22-AOAS1649

TIME-DISCRETIZATION APPROXIMATION ENRICHES CONTINUOUS-TIME DISCRETE-SPACE MODELS FOR ANIMAL MOVEMENT

BY JOSHUA HEWITT[1,a], ALAN E. GELFAND[1,b] AND ROBERT S. SCHICK[2,c]

[1]*Department of Statistical Science, Duke University,* [a]*joshua.hewitt@duke.edu,* [b]*alan@duke.edu*
[2]*Nicholas School of the Environment, Duke University,* [c]*robert.schick@duke.edu*

Continuous time discrete state models are a valuable tool for explaining animal movement. However, data collection to fit such models over a specified window of time can be misaligned with the actual realization of the movement process. This necessitates approximate model fitting, at present, through approximate imputation distributions (AIDs). Here, we propose a direct time-discretization approximation to the likelihood. The approach employs familiar ideas from hidden Markov modeling. Computation is implemented through the induced infinitesimal generator matrix. Linearization of this matrix expedites computation time. Through simulation and a real data application involving whale movement, we demonstrate that this model fitting strategy can outperform AID approaches.

REFERENCES

AL-MOHY, A. H. and HIGHAM, N. J. (2009). A new scaling and squaring algorithm for the matrix exponential. *SIAM J. Matrix Anal. Appl.* **31** 970–989. MR2538661 https://doi.org/10.1137/09074721X

ALBERTSEN, C. M. (2019). Generalizing the first-difference correlated random walk for marine animal movement data. *Sci. Rep.* **9** 4017. https://doi.org/10.1038/s41598-019-40405-z

ALBERTSEN, C. M., STØTTRUP, J. G., NIELSEN, A. and CHRISTOFFERSEN, M. O. (2014). Depth preference in released juvenile turbot Psetta maxima. *Journal of Experimental Marine Biology and Ecology* **461** 179–184.

ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* **18** 343–373. MR2461882 https://doi.org/10.1007/s11222-008-9110-y

BARLOW, J., SCHORR, G. S., FALCONE, E. A. and MORETTI, D. (2020). Variation in dive behavior of Cuvier's beaked whales with seafloor depth, time-of-day, and lunar illumination. *Mar. Ecol. Prog. Ser.* **644** 199–214.

BECKER, J., SANDWELL, D., SMITH, W., BRAUD, J., BINDER, B., DEPNER, J., FABRE, D., FACTOR, J., INGALLS, S. et al. (2009). Global bathymetry and elevation data at 30 arc seconds resolution: SRTM30_PLUS. *Marine Geodesy* **32** 355–371.

BRENNAN, A., HANKS, E. M., MERKLE, J. A., COLE, E. K., DEWEY, S. R., COURTEMANCH, A. B. and CROSS, P. C. (2018). Examining speed versus selection in connectivity models using elk migration as an example. *Landsc. Ecol.* **33** 955–968.

BRILLINGER, D. R. (2003). Simulating constrained animal motion using stochastic differential equations. In *Probability, Statistics and Their Applications: Papers in Honor of Rabi Bhattacharya. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **41** 35–48. IMS, Beachwood, OH. MR1999413 https://doi.org/10.1214/lnms/1215091656

BUDERMAN, F. E., HOOTEN, M. B., ALLDREDGE, M. W., HANKS, E. M. and IVAN, J. S. (2018). Time-varying predatory behavior is primary predictor of fine-scale movement of wildland-urban cougars. *Mov. Ecol.* **6** 1–16.

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

DOYTCHINOV, B. and IRBY, R. (2010). Time discretization of Markov chains. *Pi Mu Epsilon Journal* **13** 69–82.

GRAMACY, R. B. (2020). *Surrogates—Gaussian Process Modeling, Design, and Optimization for the Applied Sciences. Chapman & Hall/CRC Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR4283556 https://doi.org/10.1201/9780367815493

HAMILTON, J. D. (1994). *Time Series Analysis*. Princeton Univ. Press, Princeton, NJ. MR1278033

Key words and phrases. Directional persistence, infinitesimal generator, measurement error, posterior approximation, hidden Markov modeling.

HANKS, E. M., HOOTEN, M. B. and ALLDREDGE, M. W. (2015). Continuous-time discrete-space models for animal movement. *Ann. Appl. Stat.* **9** 145–165. MR3341111 https://doi.org/10.1214/14-AOAS803

HANKS, E. M., JOHNSON, D. S. and HOOTEN, M. B. (2017). Reflected stochastic differential equation models for constrained animal movement. *J. Agric. Biol. Environ. Stat.* **22** 353–372. MR3692469 https://doi.org/10.1007/s13253-017-0291-8

HARRIS, C. M., THOMAS, L., FALCONE, E. A., HILDEBRAND, J., HOUSER, D., KVADSHEIM, P. H., LAM, F.-P. A., MILLER, P. J. O., MORETTI, D. J. et al. (2018). Marine mammals and sonar: Dose-response studies, the risk-disturbance hypothesis and the role of exposure context. *J. Appl. Ecol.* **55** 396–404.

HEWITT, J., GELFAND, A. E. and SCHICK, R. S. (2023). Supplement to "Time-discretization approximation enriches continuous-time discrete-space models for animal movement." https://doi.org/10.1214/22-AOAS1649SUPPA, https://doi.org/10.1214/22-AOAS1649SUPPB

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.

LOPEZ, R. and MALARDÉ, J. (2011). Improving ARGOS Doppler location using Kalman filtering. CLS Group CLS–DT–MEMO–11–65.

LOPEZ, R., MALARDÉ, J.-P., ROYER, F. and GASPAR, P. (2013). Improving Argos Doppler location using multiple-model Kalman filtering. *IEEE Trans. Geosci. Remote Sens.* **52** 4744–4755.

MCCLINTOCK, B. T., LONDON, J. M., CAMERON, M. F. and BOVENG, P. L. (2015). Modelling animal movement using the Argos satellite telemetry location error ellipse. *Methods Ecol. Evol.* **6** 266–277.

MEWS, S., LANGROCK, R., ÖTTING, M., YAQINE, H. and REINECKE, J. (2022). Maximum approximate likelihood estimation of general continuous-time state-space models. *Stat. Model.*.

MICHELOT, T., GLENNIE, R., HARRIS, C. and THOMAS, L. (2021). Varying-coefficient stochastic differential equations with applications in ecology. *J. Agric. Biol. Environ. Stat.* **26** 446–463. MR4292797 https://doi.org/10.1007/s13253-021-00450-6

MICHELOT, T., GLOAGUEN, P., BLACKWELL, P. G. and ÉTIENNE, M.-P. (2019). The Langevin diffusion as a continuous-time model of animal movement and habitat selection. *Methods Ecol. Evol.* **10** 1894–1907.

PHILLIPS, R., SILK, J., CROXALL, J., AFANASYEV, V. and BRIGGS, D. (2004). Accuracy of geolocation estimates for flying seabirds. *Mar. Ecol. Prog. Ser.* **266** 265–272.

PREISLER, H. K., AGER, A. A. and WISDOM, M. J. (2013). Analyzing animal movement patterns using potential functions. *Ecosphere* **4** 1–13.

QUICK, N. J., CIOFFI, W. R., SHEARER, J. and READ, A. J. (2019). Mind the gap—optimizing satellite tag settings for time series analysis of foraging dives in Cuvier's beaked whales (ziphius cavirostris). *Anim. Biotelem.* **7** 5.

RAO, V. and TEH, Y. W. (2013). Fast MCMC sampling for Markov jump processes and extensions. *J. Mach. Learn. Res.* **14** 3295–3320. MR3144463

ROYER, F., FROMENTIN, J.-M. and GASPAR, P. (2005). A state–space model to derive bluefin tuna movement and habitat from archival tags. *Oikos* **109** 473–484.

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

RUSSELL, J. C., HANKS, E. M., HARAN, M. and HUGHES, D. (2018). A spatially varying stochastic differential equation model for animal movement. *Ann. Appl. Stat.* **12** 1312–1331. MR3834305 https://doi.org/10.1214/17-AOAS1113

SANTOS, M. B., PIERCE, G. J., HERMAN, J., LÓPEZ, A., GUERRA, A., MENTE, E. and CLARKE, M. (2001). Feeding ecology of Cuvier's beaked whale (Ziphius cavirostris): A review with new information on the diet of this species. *Journal of the Marine Biological Association of the United Kingdom* **81** 687–694.

SCHARF, H., HOOTEN, M. B. and JOHNSON, D. S. (2017). Imputation approaches for animal movement modeling. *J. Agric. Biol. Environ. Stat.* **22** 335–352. MR3692468 https://doi.org/10.1007/s13253-017-0294-5

SCHICK, R. S., BOWERS, M., DERUITER, S., FRIEDLAENDER, A., JOSEPH, J., MARGOLINA, T., NOWACEK, D. P. and SOUTHAL, B. L. (2019). Accounting for positional uncertainty when modeling received levels for tagged cetaceans exposed to sonar. *Aquatic Mammals* **45** 675–690.

SHEARER, J. M., QUICK, N. J., CIOFFI, W. R., BAIRD, R. W., WEBSTER, D. L., FOLEY, H. J., SWAIM, Z. T., WAPLES, D. M., BELL, J. T. et al. (2019). Diving behaviour of Cuvier's beaked whales (. *R. Soc. Open Sci.* **6** 181728. https://doi.org/10.1098/rsos.181728

SOUTHALL, B. L., BENOIT-BIRD, K. J., MOLINE, M. A. and MORETTI, D. (2019). Quantifying deep-sea predator–prey dynamics: Implications of biological heterogeneity for beaked whale conservation. *J. Appl. Ecol.* **56** 1040–1049.

THORSON, J. T., BARBEAUX, S. J., GOETHEL, D. R., KEARNEY, K. A., LAMAN, E. A., NIELSEN, J. K., SISKEY, M. R., SIWICKE, K. and THOMPSON, G. G. (2021). Estimating fine-scale movement rates and habitat preferences using multiple data sources. *Fish and Fisheries* **22** 1359–1376.

THYGESEN, U. H., PEDERSEN, M. W. and MADSEN, H. (2009). Geolocating fish using hidden Markov models and data storage tags. In *Tagging and Tracking of Marine Animals with Electronic Devices* 277–293. Springer, Berlin.

TYACK, P. L., JOHNSON, M., SOTO, N. A., STURLESE, A. and MADSEN, P. T. (2006). Extreme diving of beaked whales. *Journal of Experimental Biology* **209** 4238–4253.

WENSVEEN, P. J., ISOJUNNO, S., HANSEN, R. R., VON BENDA-BECKMANN, A. M., KLEIVANE, L., VAN IJSSELMUIDE, S., LAM FRANS-PETER, A., KVADSHEIM, P. H., DERUITER, S. L. et al. (2019). Northern bottlenose whales in a pristine environment respond strongly to close and distant navy sonar signals. *Proc. R. Soc. Lond., B Biol. Sci.* **286** 20182592.

YIN, G. G. and ZHANG, Q. (2013). *Continuous-Time Markov Chains and Applications: A Two-Time-Scale Approach*, 2nd ed. *Stochastic Modelling and Applied Probability* **37**. Springer, New York. MR2985157 https://doi.org/10.1007/978-1-4614-4346-9

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 761–785
https://doi.org/10.1214/22-AOAS1650

NESTED CONFORMAL PREDICTION SETS FOR CLASSIFICATION WITH APPLICATIONS TO PROBATION DATA

BY ARUN K. KUCHIBHOTLA[1,b] AND RICHARD A. BERK[2,d]

[1]*Department of Statistics and Data Science, Carnegie Mellon University,* [a]*arunku@cmu.edu,* [b]*https://arun-kuchibhotla.github.io*

[2]*Department of Criminology, Department of Statistics, University of Pennsylvania,* [c]*berkr@sas.upenn.edu,* [d]*https://crim.sas.upenn.edu/people/richard-berk*

Risk assessments to help inform criminal justice decisions have been used in the United States since the 1920s. Over the past several years, statistical learning risk algorithms have been introduced amid much controversy about fairness, transparency and accuracy. In this paper we focus on accuracy for a large department of probation and parole that is considering a major revision of its current, statistical learning risk methods. Because the content of each offender's supervision is substantially shaped by a forecast of subsequent conduct, forecasts have real consequences. Here, we consider the probability that risk forecasts are correct. We augment standard statistical learning estimates of forecasting uncertainty (i.e., confusion tables) with uncertainty estimates from nested conformal prediction sets. In a demonstration of concept using data from the department of probation and parole, we show that the standard uncertainty measures and uncertainty measures from nested conformal prediction sets can differ dramatically in formulation and output. We also provide a modification of nested conformal, called the localized conformal method, to match confusion tables more closely. A strong case can be made favoring the nested conformal approach. As best, we can tell our formulation of such comparisons and consequent recommendations is novel.

REFERENCES

AHLMAN, L. C. and KURTZ, E. M. (2008). The Appd Randomized Controlled Trial in Low Risk Supervision. Technical report, Internal report, Philadelphia, PA, First Judicial District of Pennsylvania.

ALEMANY, S., BELTRAN, J., PEREZ, A. and GANZFRIED, S. (2019). Predicting hurricane trajectories using a recurrent neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence* **33** 468–475.

ANGELOPOULOS, A., BATES, S., MALIK, J. and JORDAN, M. I. (2020). Uncertainty sets for image classifiers using conformal prediction. Available at arXiv:2009.14193.

BARBER, R. F., CANDÈS, E. J., RAMDAS, A. and TIBSHIRANI, R. J. (2021a). The limits of distribution-free conditional predictive inference. *Inf. Inference* **10** 455–482. MR4270755 https://doi.org/10.1093/imaiai/iaaa017

BARBER, R. F., CANDES, E. J. RAMDAS, A. and TIBSHIRANI, R. J. (2021b). Predictive inference with the jackknife+. *Ann. Statist.* **49** 486–507. MR4206687 https://doi.org/10.1214/20-AOS1965

BARNES, G. C., AHLMAN, L., GILL, C., SHERMAN, L. W., KURTZ, E. and MALVESTUTO, R. (2010). Low-intensity community supervision for low-risk offenders: A randomized, controlled trial. *Journal of Experimental Criminology* **6** 159–189.

BERK, R. (2009). The role of race in forecasts of violent crime. *Race and Social Problems* **1** 231.

BERK, R. (2019). *Machine Learning Risk Assessments in Criminal Justice Settings*. Springer, Berlin.

BERK, R. A. (2020). *Statistical Learning from a Regression Perspective*. *Springer Texts in Statistics*. Springer, Cham. MR4240261 https://doi.org/10.1007/978-3-030-40189-4

BERK, R. and KUCHIBHOTLA, A. K. (2020). Improving fairness in criminal justice algorithmic risk assessments using conformal prediction sets. Available at arXiv:2008.11664.

BERK, R., BARNES, G., AHLMAN, L. and KURTZ, E. (2010). When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment. *Journal of Experimental Criminology* **6** 191–208.

Key words and phrases. Conformal prediction, level sets, classification, risk assessment, criminal justice, statistical learning.

BERK, R., OLSON, M., BUJA, A. and OUSS, A. (2020). Using recursive partitioning to find and estimate heterogenous treatment effects in randomized clinical trials. *Journal of Experimental Criminology* 1–20.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BURGESS, E. W. (1928). Factors determining success or failure on parole. *The Workings of the Indeterminate Sentence Law and the Parole System in Illinois* 221–234.

CARLSON, A. M. (2017). The need for transparency in the age of predictive sentencing algorithms. *Iowa L. Rev.* **103** 303.

CHEN, S., GUO, Z. and ZHAO, X. (2021). Predicting mortgage early delinquency with machine learning methods. *European J. Oper. Res.* **290** 358–372. MR4188963 https://doi.org/10.1016/j.ejor.2020.07.058

CHEN, T. and GUESTRIN, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the* 22*nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining* 785–794.

COGLIANESE, C. and LEHR, D. (2019). Transparency and algorithmic governance. *Admin. L. Rev.* **71**.

GUAN, L. and TIBSHIRANI, R. (2022). Prediction and outlier detection in classification problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 524–546. MR4412996 https://doi.org/10.1111/rssb.12443

GUPTA, C., KUCHIBHOTLA, A. K. and RAMDAS, A. K. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition* **127** 108496.

HECHTLINGER, Y., PÓCZOS, B. and WASSERMAN, L. (2018). Cautious deep learning. Available at arXiv:1805.09460.

HUQ, A. Z. (2018). Racial equity in algorithmic criminal justice. *Duke LJ* **68** 1043.

IZBICKI, R., SHIMIZU, G. T. and STERN, R. B. (2019). Flexible distribution-free conditional predictive bands using density estimators. Available at arXiv:1910.05575.

KEARNS, M. and ROTH, A. (2019). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford Univ. Press, London.

KIM, B., XU, C. and BARBER, R. F. (2020). Predictive inference is free with the jackknife+-after-bootstrap. Available at arXiv:2002.09025.

LEI, J. and WASSERMAN, L. (2014). Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 71–96. MR3153934 https://doi.org/10.1111/rssb.12021

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

PAPADOPOULOS, H., PROEDROU, K., VOVK, V. and GAMMERMAN, A. (2002). Inductive confidence machines for regression. In *Machine Learning: ECML* 2002. *Lecture Notes in Computer Science* **2430** 345–356. Springer, Berlin. MR2050303 https://doi.org/10.1007/3-540-36755-1_29

ROMANO, Y., BARBER, R. F., SABATTI, C. and CANDÈS, E. J. (2019). With malice towards none: Assessing uncertainty via equalized coverage. Available at arXiv:1908.05428.

ROMANO, Y., PATTERSON, E. and CANDES, E. (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems* 3543–3553.

ROMANO, Y., SESIA, M. and CANDÈS, E. J. (2020). Classification with valid and adaptive coverage. Available at arXiv:2006.02544.

RUDIN, C., WANG, C. and COKER, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review* **2**.

SADINLE, M., LEI, J. and WASSERMAN, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *J. Amer. Statist. Assoc.* **114** 223–234. MR3941250 https://doi.org/10.1080/01621459.2017.1395341

VOVK, V. (2013). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, 475–490. PMLR.

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

YAN, C., YAO, J., LI, R., XU, Z. and HUANG, J. (2018). Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays. In *Proceedings of the* 2018 *ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* 103–110.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 786–814
https://doi.org/10.1214/22-AOAS1651

HANDLING CATEGORICAL FEATURES WITH MANY LEVELS USING A PRODUCT PARTITION MODEL

BY TULIO L. CRISCUOLO[1,a], RENATO M. ASSUNÇÃO[2,c], ROSANGELA H. LOSCHI[3,d], WAGNER MEIRA JR.[1,b] AND DANNA CRUZ-REYES[4,e]

[1]*Departamento de Ciência da Computação, Universidade Federal de Minas Gerais,* [a]*tcriscuolo@dcc.ufmg.br,* [b]*meira@dcc.ufmg.br*

[2]*ESRI Inc.,* [c]*rassuncao@esri.com*

[3]*Departmento de Estatística, Universidade Federal de Minas Gerais,* [d]*loschi@est.ugmg.br*

[4]*Grupo de Investigación Clínica, Escuela de Medicina y Ciencias de la Salud, Universidad del Rosario,* [e]*dcruzreyes@gmail.com*

A common difficulty in data analysis is how to handle categorical predictors with a large number of levels or categories. Few proposals have been developed to tackle this important and frequent problem. We introduce a generative model that simultaneously carries out the model fitting and the aggregation of the categorical levels into larger groups. We represent the categorical predictor by a graph where the nodes are the categories and establish a probability distribution over meaningful partitions of this graph. Conditionally on the observed data, we obtain a posterior distribution for the levels aggregation, allowing the inference about the most probable clustering for the categories. Simultaneously, we extract inference about all the other regression model parameters. We compare our and state-of-art methods showing that it has equally good predictive performance and more interpretable results. Our approach balances out accuracy vs. interpretability, a current important concern in statistics and machine learning.

REFERENCES

BONDELL, H. D. and REICH, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics* **65** 169–177. MR2665858 https://doi.org/10.1111/j.1541-0420.2008.01061.x

CRISCUOLO, T. L, ASSUNÇÃO, R. M, LOSCHI, R. H, MEIRA JR., W. and CRUZ-REYES, D. (2023). Supplement to "Handling categorical features with many levels using a product partition model." https://doi.org/10.1214/22-AOAS1651SUPP

CRITEOLABS (2014). Display Advertising Challenge. Available at https://www.kaggle.com/c/criteo-display-ad-challenge.

CROSS-VALIDATED (2013). How to handle categorical predictors with too many levels? Available at https://stats.stackexchange.com/questions/67938/how-to-handle-categorical-predictors-with-too-many-levels.

CROSS-VALIDATED (2017). Principled way of collapsing categorical variables with many levels? Available at https://stats.stackexchange.com/questions/146907/principled-way-of-collapsing-categorical-variables-with-many-levels.

DELLAPORTAS, P. and TARANTOLA, C. (2005). Model determination for categorical data with factor level merging. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 269–283. MR2137325 https://doi.org/10.1111/j.1467-9868.2005.00501.x

GERTHEISS, J. and TUTZ, G. (2010). Sparse modeling of categorial explanatory variables. *Ann. Appl. Stat.* **4** 2150–2180. MR2829951 https://doi.org/10.1214/10-AOAS355

GUSTAFSON, P. (2005). On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statist. Sci.* **20** 111–140. MR2183445 https://doi.org/10.1214/088342305000000098

HALLAC, D., LESKOVEC, J. and BOYD, S. (2015). Network lasso: Clustering and optimization in large graphs. In *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 387–396. ACM, New York.

Key words and phrases. Categorical predictors, clustering effects, random partition, dimension reduction, linear regression.

HARTIGAN, J. A. (1990). Partition models. *Comm. Statist. Theory Methods* **19** 2745–2756. MR1088047 https://doi.org/10.1080/03610929008830345

HEGARTY, A. and BARRY, D. (2008). Bayesian disease mapping using product partition models. *Stat. Med.* **27** 3868–3893. MR2526613 https://doi.org/10.1002/sim.3253

LEYS, C., LEY, C., KLEIN, O., BERNARD, P. and LICATA, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* **49** 764–766.

LUO, Z. T., SANG, H. and MALLICK, B. (2021). A Bayesian contiguous partitioning method for learning clustered latent variables. *J. Mach. Learn. Res.* **22** Paper No. 37. MR4253730

MCGINNIS, W. (2019). Category encoders. Available at http://contrib.scikit-learn.org/categorical-encoding/index.html.

MICCI-BARRECA, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explor. Newsl.* **3** 27–32.

MOUNTAND, J. and ZUMEL, N. (2019). vtreat: A statistically sound 'data.frame' processor/conditioner. Available at https://cran.r-project.org/web/packages/vtreat/index.html.

OELKER, M.-R. (2015). gvcm.cat: regularized categorical effects/categorical effect modifiers/continuous/smooth effects in GLMs. R package version 1.9.

OELKER, M.-R., GERTHEISS, J. and TUTZ, G. (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Stat. Model.* **14** 157–177. MR3192557 https://doi.org/10.1177/1471082X13503452

OLIVEIRA, L. G., ARGIENTO, R., LOSCHI, R. H., ASSUNÇÃO, R. M., RUGGERI, F. and BRANCO, M. D. (2022). Bias correction in clustered underreported data. *Bayesian Anal.* **17** 95–126. MR4377138 https://doi.org/10.1214/20-ba1244

PAGE, G. L. and QUINTANA, F. A. (2016). Spatial product partition models. *Bayesian Anal.* **11** 265–298. MR3465813 https://doi.org/10.1214/15-BA971

PAUGER, D. and WAGNER, H. (2019). Bayesian effect fusion for categorical predictors. *Bayesian Anal.* **14** 341–369. MR3934089 https://doi.org/10.1214/18-BA1096

PAUGER, D., LEITNER, M., WAGNER, H. and MALSINER-WALLI, G. (2019). effectFusion: Bayesian effect fusion for categorical predictors. R package version 1.1.1. Available at https://CRAN.R-project.org/package=effectFusion.

PROKHORENKOVA, L., GUSEV, G., VOROBEV, A., DOROGUSH, A. V. and GULIN, A. (2017). CatBoost: Unbiased boosting with categorical features. R package version 0.14.2.

PROKHORENKOVA, L., GUSEV, G., VOROBEV, A., DOROGUSH, A. V. and GULIN, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems* 6638–6648.

ROSTAMIZADEH, A., ESFANDIARI, H., CHEN, L., BATENI, M., FU, T. and MIRROKNI, V. (2019). Categorical feature compression via submodular optimization. In *International Conference on Machine Learning* 515–523.

SCHAUBERGER, G. and TUTZ, G. (2014). catdata: Categorical data. R-package. Available at https://CRAN.R-project.org/package=catdata.

TEIXEIRA, L. V., ASSUNÇÃO, R. M. and LOSCHI, R. H. (2015). A generative spatial clustering model for random data through spanning trees. In 2015 *IEEE International Conference on Data Mining* 997–1002.

TEIXEIRA, L. V., ASSUNÇÃO, R. M. and LOSCHI, R. H. (2019). Bayesian space–time partitioning by sampling and pruning spanning trees. *J. Mach. Learn. Res.* **20** Paper No. 85. MR3960939

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

TUTZ, G. and BERGER, M. (2018). Tree-structured modelling of categorical predictors in generalized additive regression. *Adv. Data Anal. Classif.* **12** 737–758. MR3854425 https://doi.org/10.1007/s11634-017-0298-6

WIN-VECTOR (2012). Modeling trick: Impact coding of categorical variables with many levels. Available at http://www.win-vector.com/blog/2012/07/modeling-trick-impact-coding-of-categorical-variables-with-many-levels/.

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

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 815–837
https://doi.org/10.1214/22-AOAS1652

PALM: PATIENT-CENTERED TREATMENT RANKING VIA LARGE-SCALE MULTIVARIATE NETWORK META-ANALYSIS

BY RUI DUAN[1,a], JIAYI TONG[2,b], LIFENG LIN[3,d], LISA LEVINE[4,e], MARY SAMMEL[5,f], JOEL STODDARD[5,g], TIANJING LI[5,h], CHRISTOPHER H SCHMID[6,i], HAITAO CHU[7,j] AND YONG CHEN[2,c]

[1]*Department of Biostatistics, Harvard T.H. Chan School of Public Health,* [a]*rduan@hsph.harvard.edu*

[2]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania,* [b]*jiayi.tong@pennmedicine.upenn.edu,* [c]*ychen123@upenn.edu*

[3]*Department of Epidemiology and Biostatistics, University of Arizona,* [d]*lifenglin@arizona.edu*

[4]*Department of Obstetrics and Gynecology, University of Pennsylvania,* [e]*lisa.levine@uphs.upenn.edu*

[5]*University of Colorado Denver,* [f]*mary.sammel@cuanschutz.edu,* [g]*joel.stoddard@ucdenver.edu,* [h]*tianjing.li@cuanschutz.edu*

[6]*Department of Biostatistics, Brown University,* [i]*christopher_schmid@brown.edu*

[7]*Statistical Research and Data Science Center, Pfizer Inc.,* [j]*Haitao.Chu@Pfizer.com*

The growing number of available treatment options has led to urgent needs for reliable answers when choosing the best course of treatment for a patient. As it is often infeasible to compare a large number of treatments in a single randomized controlled trial, multivariate network meta-analyses (NMAs) are used to synthesize evidence from trials of a subset of the treatments, where both efficacy and safety related outcomes are considered simultaneously. However, these large-scale multiple-outcome NMAs have created challenges to existing methods due to the increasing complexity of the unknown correlations between outcomes and treatment comparisons. In this paper, we proposed a new framework for PAtient-centered treatment ranking via Large-scale Multivariate network meta-analysis, termed as PALM, which includes a parsimonious modeling approach, a fast algorithm for parameter estimation and inference, a novel visualization tool for presenting multivariate outcomes, termed as the origami plot, as well as personalized treatment ranking procedures taking into account the individual's considerations on multiple outcomes. In application to an NMA that compares 14 treatment options for labor induction, we provided a comprehensive illustration of the proposed framework and demonstrated its computational efficiency and practicality, and we obtained new insights and evidence to support patient-centered clinical decision making.

REFERENCES

ALFIREVIC, Z., KEENEY, E., DOWSWELL, T. et al. (2015). Labour induction with prostaglandins: A systematic review and network meta-analysis. *BMJ* **350** h217.

BOULVAIN, M., KELLY, A. J. and IRION, O. (2008). Intracervical prostaglandins for induction of labour. *Cochrane Database of Systematic Reviews* **1** Art. No.: CD006971.

BYGDEMAN, M. (2003). Pharmacokinetics of prostaglandins. *Best Pract. Res., Clin. Obstet. Gynaecol.* **17** 707–716.

CALDWELL, D. M., ADES, A. E. and HIGGINS, J. P. T. (2005). Simultaneous comparison of multiple treatments: Combining direct and indirect evidence. *BMJ* **331** 897–900.

CAUGHEY, A. B., CAHILL, A. G., GUISE, J.-M. and ROUSE, D. J. (2014). Safe prevention of the primary cesarean delivery. *Am. J. Obstet. Gynecol.* **210** 179–193.

CHAIMANI, A., HIGGINS, J. P. T., MAVRIDIS, D., SPYRIDONOS, P. and SALANTI, G. (2013). Graphical tools for network meta-analysis in STATA. *PLoS ONE* **8** e76654.

CHEN, Y., FAN, J., MA, C. and WANG, K. (2019). Spectral method and regularized MLE are both optimal for top-K ranking. *Ann. Statist.* **47** 2204–2235. MR3953449 https://doi.org/10.1214/18-AOS1745

Key words and phrases. Evidence-based medicine, multiple outcomes, network meta-analysis, origami plot, treatment ranking.

CHEN, Y., HONG, C. and RILEY, R. D. (2015). An alternative pseudolikelihood method for multivariate random-effects meta-analysis. *Stat. Med.* **34** 361–380. MR3301582 https://doi.org/10.1002/sim.6350

CHEN, Y., LIU, Y., CHU, H., LEE, M.-L. T. and SCHMID, C. H. (2017b). A simple and robust method for multivariate meta-analysis of diagnostic test accuracy. *Stat. Med.* **36** 105–121. MR3580948 https://doi.org/10.1002/sim.7093

CHEN, Y., LIU, Y., NING, J., CORMIER, J. and CHU, H. (2015). A hybrid model for combining case-control and cohort studies in systematic reviews of diagnostic tests. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 469–489. MR3325459 https://doi.org/10.1111/rssc.12087

CHEN, Y., LIU, Y., NING, J., NIE, L., ZHU, H. and CHU, H. (2017a). A composite likelihood method for bivariate meta-analysis in diagnostic systematic reviews. *Stat. Methods Med. Res.* **26** 914–930. MR3635944 https://doi.org/10.1177/0962280214562146

CHEN, Y., NING, J. and CAI, C. (2015). Regression analysis of longitudinal data with irregular and informative observation times. *Biostatistics* **16** 727–739. MR3449839 https://doi.org/10.1093/biostatistics/kxv008

CIPRIANI, A., BARBUI, C. et al. (2011). Comparative efficacy and acceptability of antimanic drugs in acute mania: A multiple-treatments meta-analysis. *Lancet* **378** 1306–1315.

CIPRIANI, A., FURUKAWA, T. A. et al. (2018). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: A systematic review and network meta-analysis. *Lancet* **391** 1357–1366.

COPAS, J. and SHI, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics* **1** 247–262.

COX, D. R. and REID, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91** 729–737. MR2090633 https://doi.org/10.1093/biomet/91.3.729

DAVEY, J., TURNER, R. M., CLARKE, M. J. and HIGGINS, J. P. T. (2011). Characteristics of meta-analyses and their component studies in the *cochrane database of systematic reviews*: A cross-sectional, descriptive analysis. *BMC Med. Res. Methodol.* **11** 160.

DIAS, S., WELTON, N. J., SUTTON, A. J. et al. (2013). Evidence synthesis for decision making 4: Inconsistency in networks of evidence based on randomized controlled trials. *Med. Decis. Mak.* **33** 641–656.

DUAN, R., NING, Y., WANG, S., LINDSAY, B. G., CARROLL, R. J. and CHEN, Y. (2020). A fast score test for generalized mixture models. *Biometrics* **76** 811–820. MR4151850 https://doi.org/10.1111/biom.13204

DUAN, R., TONG, J., LIN, L., LEVINE, L., SAMMEL, M., STODDARD, J., LI, T., SCHMID, C. H., CHU, H. and CHEN, Y. (2023). Supplement to "PALM: Patient-centered treatment ranking via large-scale multivariate network meta-analysis." https://doi.org/10.1214/22-AOAS1652SUPP

DULAI, P. S., SINGH, S. et al. (2016). Chemoprevention of colorectal cancer in individuals with previous colorectal neoplasia: Systematic review and network meta-analysis. *BMJ* **355** i6188.

DUNKLEY, A., CHARLES, K., GRAY, L. et al. (2012). Effectiveness of interventions for reducing diabetes and cardiovascular disease risk in people with metabolic syndrome: Systematic review and mixed treatment comparison meta-analysis. *Diabetes Obes. Metab.* **14** 616–625.

EFTHIMIOU, O., MAVRIDIS, D., CIPRIANI, A., LEUCHT, S., BAGOS, P. and SALANTI, G. (2014). An approach for modelling multiple correlated outcomes in a network of interventions using odds ratios. *Stat. Med.* **33** 2275–2287. MR3256692 https://doi.org/10.1002/sim.6117

EFTHIMIOU, O., MAVRIDIS, D., RILEY, R. D., CIPRIANI, A. and SALANTI, G. (2015). Joint synthesis of multiple correlated outcomes in networks of interventions. *Biostatistics* **16** 84–97. MR3365413 https://doi.org/10.1093/biostatistics/kxu030

CENTERS FOR DISEASE CONTROL AND PREVENTION (2020) (accessed February 3, 2020). National Center for Health Statistics–Births.

GUYATT, G. H., OXMAN, A. D., KUNZ, R. et al. (2011). GRADE guidelines: 8. Rating the quality of evidence—indirectness. *J. Clin. Epidemiol.* **64** 1303–1310.

HONG, C., NING, Y., WANG, S., WU, H., CARROLL, R. J. and CHEN, Y. (2017). PLEMT: A novel pseudolikelihood-based EM test for homogeneity in generalized exponential tilt mixture models. *J. Amer. Statist. Assoc.* **112** 1393–1404. MR3750863 https://doi.org/10.1080/01621459.2017.1280405

HONG, C., SALANTI, G., MORTON, S. C., RILEY, R. D., CHU, H., KIMMEL, S. E. and CHEN, Y. (2020). Testing small study effects in multivariate meta-analysis. *Biometrics* **76** 1240–1250. MR4186839 https://doi.org/10.1111/biom.13342

JACKSON, D., BUJKIEWICZ, S., LAW, M., RILEY, R. D. and WHITE, I. R. (2018). A matrix-based method of moments for fitting multivariate network meta-analysis models with multiple outcomes and random inconsistency effects. *Biometrics* **74** 548–556. MR3825341 https://doi.org/10.1111/biom.12762

JACKSON, D., RILEY, R. and WHITE, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Stat. Med.* **30** 2481–2498. MR2843472 https://doi.org/10.1002/sim.4172

KIRKHAM, J. J., RILEY, R. D. and WILLIAMSON, P. R. (2012). A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Stat. Med.* **31** 2179–2195. MR2967921 https://doi.org/10.1002/sim.5356

LI, F., HUANG, W. and ZHANG, X. (2018). Efficacy and safety of different regimens for primary open-angle glaucoma or ocular hypertension: A systematic review and network meta-analysis. *Acta Ophthalmol*. **96** e277–e284. https://doi.org/10.1111/aos.13568

LI, T., LINDSLEY, K., ROUSE, B. et al. (2016). Comparative effectiveness of first-line medications for primary open-angle glaucoma: A systematic review and network meta-analysis. *Ophthalmology* **123** 129–140.

LIANG, K.-Y. (1983). On information and ancillarity in the presence of a nuisance parameter. *Biometrika* **70** 607–612. MR0725374 https://doi.org/10.1093/biomet/70.3.607

LIANG, K.-Y. and ZEGER, S. L. (1995). Inference based on estimating functions in the presence of nuisance parameters. *Statist*. *Sci*. **10** 158–173. MR1368098

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** 5.

LINDSAY, B. G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes* (*Ithaca*, *NY*, 1987). *Contemp*. *Math*. **80** 221–239. Amer. Math. Soc., Providence, RI. MR0999014 https://doi.org/10.1090/conm/080/999014

LIU, Y., DESANTIS, S. M. and CHEN, Y. (2018). Bayesian mixed treatment comparisons meta-analysis for correlated outcomes subject to reporting bias. *J*. *R*. *Stat*. *Soc*. *Ser*. *C*. *Appl*. *Stat*. **67** 127–144. MR3758758 https://doi.org/10.1111/rssc.12220

LIU, Y., FANG, E. X. and LU, J. (2022). Lagrangian inference for ranking problems. *Oper*. *Res*. https://doi.org/10.1287/opre.2022.2313

LU, G. and ADES, A. E. (2006). 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.

LUMLEY, T. (2002). Network meta-analysis for indirect treatment comparisons. *Stat*. *Med*. **21** 2313–2324.

LUO, C., MARKS-ANGLIN, A., DUAN, R., LIN, L., HONG, C., CHU, H. and CHEN, Y. (2022). Accounting for publication bias using a bivariate trim and fill meta-analysis procedure. *Stat*. *Med*. **41** 3466–3478. MR4453458 https://doi.org/10.1002/sim.9428

MARKS-ANGLIN, A. and CHEN, Y. (2020a). A historical review of publication bias. *Res*. *Synth*. *Methods* **11** 725–742.

MARKS-ANGLIN, A. and CHEN, Y. (2020b). Small-study effects: Current practice and challenges for future research. *Stat*. *Interface* **13** 475–484. MR4131724 https://doi.org/10.4310/SII.2020.v13.n4.a5

MARKS-ANGLIN, A., DUAN, R., CHEN, Y., PANAGIOTOU, O. and SCHMID, C. H. (2020). Publication and outcome reporting bias. In *Handbook of Meta-Analysis* 283–312. CRC Press/CRC, Boca Raton.

MARKS-ANGLIN, A., LUO, C., PIAO, J., GIBBONS, M. B. C., SCHMID, C. H., NING, J. and CHEN, Y. (2022). EMBRACE: An EM-based bias reduction approach through Copas-model estimation for quantifying the evidence of selective publishing in network meta-analysis. *Biometrics* **78** 754–765. MR4450592 https://doi.org/10.1111/biom.13441

MAURI, D., POLYZOS, N. P., SALANTI, G., PAVLIDIS, N. and IOANNIDIS, J. P. (2008). Multiple-treatments meta-analysis of chemotherapy and targeted therapies in advanced breast cancer. *J*. *Natl*. *Cancer Inst*. **100** 1780–1791.

MAVRIDIS, D., PORCHER, R., NIKOLAKOPOULOU, A., SALANTI, G. and RAVAUD, P. (2020). Extensions of the probabilistic ranking metrics of competing treatments in network meta-analysis to reflect clinically important relative differences on many outcomes. *Biom*. *J*. **62** 375–385. MR4075815 https://doi.org/10.1002/bimj.201900026

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

NING, Y. and CHEN, Y. (2015). A class of pseudolikelihood ratio tests for homogeneity in exponential tilt mixture models. *Scand*. *J*. *Stat*. **42** 504–517. MR3345118 https://doi.org/10.1111/sjos.12119

NING, J., CAI, C., CHEN, Y., HUANG, X. and WANG, M.-C. (2020). Semiparametric modelling and estimation of covariate-adjusted dependence between bivariate recurrent events. *Biometrics* **76** 1229–1239. MR4186838 https://doi.org/10.1111/biom.13229

NING, J., CHEN, Y., CAI, C., HUANG, X. and WANG, M.-C. (2015). On the dependence structure of bivariate recurrent event processes: Inference and estimation. *Biometrika* **102** 345–358. MR3371008 https://doi.org/10.1093/biomet/asu073

NOMA, H., TANAKA, S., MATSUI, S., CIPRIANI, A. and FURUKAWA, T. A. (2017). Quantifying indirect evidence in network meta-analysis. *Stat*. *Med*. **36** 917–927. MR3606653 https://doi.org/10.1002/sim.7187

PETROPOULOU, M., NIKOLAKOPOULOU, A., VERONIKI, A.-A. et al. (2017). Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. *J*. *Clin*. *Epidemiol*. **82** 20–28.

PILLINGER, T., MCCUTCHEON, R. A., VANO, L. et al. (2020). Comparative effects of 18 antipsychotics on metabolic function in patients with schizophrenia, predictors of metabolic dysregulation, and association with psychopathology: A systematic review and network meta-analysis. *Lancet Psychiatry* **7** 64–77.

PUSTEJOVSKY, J. E. and TIPTON, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *J. Bus. Econom. Statist.* **36** 672–683. MR3871709 https://doi.org/10.1080/07350015.2016.1247004

RILEY, R. D., JACKSON, D., SALANTI, G., BURKE, D. L., PRICE, M., KIRKHAM, J. and WHITE, I. R. (2017). Multivariate and network meta-analysis of multiple outcomes and multiple treatments: Rationale, concepts, and examples. *BMJ* **358** j3932. https://doi.org/10.1136/bmj.j3932

RILEY, R. D., THOMPSON, J. R. et al. (2007). An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics* **9** 172–186.

ROGLIANI, P., MATERA, M. G. et al. (2019). Efficacy and cardiovascular safety profile of dual bronchodilation therapy in chronic obstructive pulmonary disease: A bidimensional comparative analysis across fixed-dose combinations. *Pulm. Pharmacol. Ther.* **59** 101841.

ROSENBERGER, K. J., DUAN, R., CHEN, Y. and LIN, L. (2021). Predictive P-score for treatment ranking in Bayesian network meta-analysis. *BMC Med. Res. Methodol.* **21** 213.

RUCKER, G., SCHWARZER, G., KRAHN, U. and KONIG, J. (2017). Network meta-analysis using frequentist methods. *R Package Version* **1**.

SAARY, M. J. (2008). Radar plots: A useful way for presenting multivariate health care data. *J. Clin. Epidemiol.* **61** 311–317.

SALANTI, G. (2012). Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: Many names, many benefits, many concerns for the next generation evidence synthesis tool. *Res. Synth. Methods* **3** 80–97.

SALANTI, G., ADES, A. and IOANNIDIS, J. P. (2011). Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: An overview and tutorial. *J. Clin. Epidemiol.* **64** 163–171.

SALANTI, G., HIGGINS, J. P. T., ADES, A. E. and IOANNIDIS, J. P. A. (2008). Evaluation of networks of randomized trials. *Stat. Methods Med. Res.* **17** 279–301. MR2521537 https://doi.org/10.1177/0962280207080643

SCHMID, C. H. (2017). Outcome reporting bias: A pervasive problem in published meta-analyses. *Am. J. Kidney Dis.* **69** 172–174. https://doi.org/10.1053/j.ajkd.2016.11.003

SHEN, W., LIU, S., CHEN, Y. and NING, J. (2019). Regression analysis of longitudinal data with outcome-dependent sampling and informative censoring. *Scand. J. Stat.* **46** 831–847. MR3994171 https://doi.org/10.1111/sjos.12373

SILVER, R. M. (2012). Implications of the first cesarean: Perinatal and future reproductive health and subsequent cesareans, placentation issues, uterine rupture risk, morbidity, and mortality. *Semin. Perinatol.* **36** 315–323.

SLEE, A., NAZARETH, I. et al. (2019). Pharmacological treatments for generalised anxiety disorder: A systematic review and network meta-analysis. *Lancet* **393** 768–777.

TANG, O. S., SCHWEER, H., SEYBERTH, H., LEE, S. W. and HO, P. C. (2002). Pharmacokinetics of different routes of administration of misoprostol. *Hum. Reprod.* **17** 332–336.

TRIKALINOS, T. A., HOAGLIN, D. C. and SCHMID, C. H. (2014). Empirical and simulation-based comparison of univariate and multivariate meta-analysis for binary outcomes.

VAN VALKENHOEF, G., LU, G., DE BROCK, B., HILLEGE, H., ADES, A. and WELTON, N. J. (2012). Automating network meta-analysis. *Res. Synth. Methods* **3** 285–299.

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

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.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 838–859
https://doi.org/10.1214/22-AOAS1653

MANY-TO-ONE INDIRECT SAMPLING WITH APPLICATION TO THE FRENCH POSTAL TRAFFIC ESTIMATION

BY ESTELLE MEDOUS[1,a], CAMELIA GOGA[2,c], ANNE RUIZ-GAZEN[1,b], JEAN-FRANÇOIS BEAUMONT[3,d], ALAIN DESSERTAINE[4,e] AND PAULINE PUECH[4,f]

[1]*Toulouse School of Economics, Université Toulouse 1 Capitole 1,* [a]*estelle.medous@tse-fr.eu,* [b]*anne.ruiz-gazen@tse-fr.eu*

[2]*Laboratoire de Mathématiques de Besançon, Université de Bourgogne Franche-Comté,* [c]*camelia.goga@univ-fcomte.fr*

[3]*Modern Statistical Methods and Data science Branch, Statistics Canada,* [d]*jean-francois.beaumont@canada.ca*

[4]*Département Conception et Pilotage des Etudes Statistiques de la Branche Service Courrier Colis, La Poste,* [e]*alain.dessertaine@laposte.fr,* [f]*pauline.puech@laposte.fr*

In social and economic surveys, it can be difficult to directly reach units of the target population, and indirect sampling is often advocated to solve this issue. In indirect sampling, the sample is drawn from a frame population that is linked to the target population, and estimation of target population parameters is typically achieved through the generalized weight share method (GWSM). This method provides a weight, for every unit of the target population, that depends on the one hand, on the sampling weights in the frame population and, on the other hand, on the link weights between the frame population and the target population. In the present study, we focus on the situation in which the units from the frame population are linked to one and only one unit from the target population (Many-to-One case). This situation is encountered at the French postal service where addresses are sampled instead of postman rounds. We aim at understanding of the impact of the link weights on the efficiency of the GWSM estimators. We derive variance expressions and optimality results for a large class of sampling designs. Moreover, we note that the Many-to-One case can lead to too many links to observe. We alleviate the problem by introducing an intermediate population and double indirect sampling. The question of the loss of precision in this situation is discussed in detail through theoretical results and simulations. These findings help to explain the loss of precision of double GWSM estimators observed recently at the French postal service.

REFERENCES

DE VITIIS, C., FALORSI, S., INGLESE, F., MASI, A., PANNUZI, N. and RUSSO, M. (2014). A methodological approach based on indirect sampling to survey the homeless population. *Riv. Stat. Uff.* **1** 9–30.

DESSERTAINE, A. and FLUTEAUX, L. (2004). Utilisation de la méthode généralisée du partage des poids dans le cadre des estimations de flux de courrier à la poste. In *Echantillonnage et Méthodes D'enquêtes*, *Science Sup* (P. Ardilly, ed.) 219–227. Dunod, Paris, France.

DEVILLE, J.-C. and LAVALLÉE, P. (2006). Indirect sampling: The foundations of the generalized weight share method. *Surv. Methodol.* **32** 165–176.

DEVILLE, J.-C. and MAUMY-BERTRAND, M. (2006). Extension of the indirect sampling method and its application to tourism. *Surv. Methodol.* **32** 177–185.

FALORSI, P. D., RIGHI, P. and LAVALLÉE, P. (2019). Cost optimal sampling for the integrated observation of different populations. *Surv. Methodol.* **45** 485–511.

HAZIZA, D. and BEAUMONT, J.-F. (2017). Construction of weights in surveys: A review. *Statist. Sci.* **32** 206–226. MR3648956 https://doi.org/10.1214/16-STS608

KALTON, G. and BRICK, J. M. (1995). Weighting schemes for household panel surveys. *Surv. Methodol.* **21** 33–34.

KIESL, H. (2016). Indirect sampling: A review of theory and recent applications. *AStA Wirtsch. Sozialstat. Arch.* **10** 289–303.

Key words and phrases. Generalized weight share method, optimal link weights, stratified sampling, variance estimation.

LARDIN-PUECH, P. (2014). Estimation du trafic de courrier distribué en France métropolitaine par sondage indirect. In *The Proceedings of the* 8ème *Colloque Francophone sur les Sondages*, *Dijon*. Available at http://paperssondages14.sfds.asso.fr/submission_100.pdf.

LAVALLÉE, P. (2007). *Indirect Sampling*. *Springer Series in Statistics*. Springer, New York. MR2344875 https://doi.org/10.1007/978-0-387-70782-2

RENDTEL, U. and HARMS, T. (2009). Weighting and calibration for household panels. In *Methodology of Longitudinal Surveys* 265–286.

XU, X. and LAVALLÉE, P. (2009). Treatments for link nonresponse in indirect sampling. *Surv. Methodol*. **35** 153–164.

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 860–880
https://doi.org/10.1214/22-AOAS1655

MODELING A NONLINEAR BIOPHYSICAL TREND FOLLOWED BY LONG-MEMORY EQUILIBRIUM WITH UNKNOWN CHANGE POINT

BY WENYU ZHANG[1,a], MARYCLARE GRIFFIN[2,c] AND DAVID S. MATTESON[1,b]

[1]*Department of Statistics and Data Science, Cornell University,* [a]*wz258@cornell.edu,* [b]*dm484@cornell.edu*
[2]*Department of Mathematics and Statistics, University of Massachusetts Amherst,* [c]*maryclaregri@umass.edu*

Measurements of many biological processes are characterized by an initial trend period followed by an equilibrium period. Scientists may wish to quantify features of the two periods as well as the timing of the change point. Specifically, we are motivated by problems in the study of electrical cell-substrate impedance sensing (ECIS) data. ECIS is a popular new technology which measures cell behavior noninvasively. Previous studies using ECIS data have found that different cell types can be classified by their equilibrium behavior. However, it can be challenging to identify when equilibrium has been reached and to quantify the relevant features of cells' equilibrium behavior. In this paper we assume that measurements during the trend period are independent deviations from a smooth nonlinear function of time, and that measurements during the equilibrium period are characterized by a simple long memory model. We propose a method to simultaneously estimate the parameters of the trend and equilibrium processes and locate the change point between the two. We find that this method performs well in simulations and in practice. When applied to ECIS data, it produces estimates of change points and measures of cell equilibrium behavior which offer improved classification of infected and uninfected cells.

REFERENCES

BAILLIE, R. T. (1996). Long memory processes and fractional integration in econometrics. *J. Econometrics* **73** 5–59. https://doi.org/10.1016/0304-4076(95)01732-1

CHEN, C. and LIU, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. *J. Amer. Statist. Assoc.* **88** 284–297.

DOORNIK, J. A. and OOMS, M. (2004). Inference and forecasting for ARFIMA models with an application to US and UK inflation. *Stud. Nonlinear Dyn. Econom.* **8**.

DUFRENOT, G., GUEGAN, D. and PEGUIN-FEISSOLLE, A. (2008). Changing-regime volatility: A fractionally integrated SETAR model. *Appl. Financ. Econ.* **18** 519–526.

GELSINGER, M., TUPPER, L. and MATTESON, D. (2020). Cell line classification using Electric Cell-substrate Impedance Sensing (ECIS). *Int. J. Biostat.* **16**.

GUSTAVSSON, R., MANDENIUS, C. F., L'́OFGREN, S., SCHEPER, T. and LINDLER, P. (2019). In situ microscopy as online tool for detecting microbial contaminations in cell culture. *J. Biotechnol.* **296** 53–60.

HASLETT, J. and RAFTERY, A. E. (1989). Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **38** 1–50.

HONG, J., KANDASAMY, K., MARIMUTHU, M., CHOI, C. S. and KIM, S. (2011). Electrical cell-substrate impedance sensing as a non-invasive tool for cancer cell study. *Analyst* **136** 237–245.

JAMES, N. A. and MATTESON, D. S. (2015). ecp: An R package for nonparametric multiple change point analysis of multivariate data. *J. Stat. Softw.* **62** 1–25.

KEESE, C. (2019). ECIS Application Webinar Series. http://www.biophysics.com/webinar.php. Accessed: 2019-04-13.

KILLICK, R., FEARNHEAD, P. and ECKLEY, I. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598.

KUAN, C.-M. (2004). Generalized least squares theory.

Key words and phrases. Applied biophysics, change-point analysis, fractionally integrated process, long memory, time series.

LOVELADY, D. C., RICHMOND, T. C., MAGGI, A. N., LO, C.-M. and RABSON, D. A. (2007). Distinguishing cancerous from noncancerous cells through analysis of electrical noise. *Phys. Rev. E, Stat. Nonlin. Soft Matter Phys.* **76** 041908. https://doi.org/10.1103/PhysRevE.76.041908

MATTESON, D. S. and JAMES, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *J. Amer. Statist. Assoc.* **109** 334–345.

NIKA, V., BABYN, P. and ZHU, H. (2014). Change detection of medical images using dictionary learning techniques and PCA. In *Medical Imaging*.

OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572. https://doi.org/10.1093/biostatistics/kxh008

OPP, D., WAFULA, B., LIM, J., HUANG, E., LO, J.-C. and LO, C.-M. (2009). Use of electric cell–substrate impedance sensing to assess in vitro cytotoxicity. *Biosens. Bioelectron.* **24** 2625–2629.

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

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, Cambridge. https://doi.org/10.1017/CBO9780511755453

SHAPIRO, S. S. and WILK, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52** 591–611.

SOWELL, F. (1992). Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *J. Econometrics* **53** 165–188.

TARANTOLA, M., MAREL, A.-K., SUNNICK, E., ADAM, H., WEGENER, J. and JANSHOFF, A. (2010). Dynamics of human cancer cell lines monitored by electrical and acoustic fluctuation analysis. *Integr. Biol.* **2** 139–50. https://doi.org/10.1039/b920815a

THOMAS, O. and CORANDER, J. (2019). Diagnosing model misspecification and performing generalized Bayes' updates via probabilistic classifiers. *Methodol.* ArXiv.

ZHANG, W., GILBERT, D. and MATTESON, D. S. (2019). ABACUS: Unsupervised multivariate change detection via Bayesian source separation. In *Proceedings of the* 2019 *SIAM International Conference on Data Mining* (*SDM*) 603–611. SIAM, Philadelphia.

ZHANG, W., GRIFFIN, M. and MATTESON, D. S. (2023). Supplement to "Modeling a nonlinear biophysical trend followed by long-memory equilibrium with unknown change point." https://doi.org/10.1214/22-AOAS1655SUPP

The Annals of Applied Statistics
2023, Vol. 17, No. 1, 881–899
https://doi.org/10.1214/22-AOAS1659

AN EXTENSION OF ESTIMATING EQUATIONS TO MODEL LONGITUDINAL MEDICAL COST TRAJECTORY WITH MEDICARE CLAIMS DATA LINKED TO SEER CANCER REGISTRY

BY SHIKUN WANG[1,a], JING NING[2,b], YING XU[3,e], YA-CHEN TINA SHIH[3,f], YU SHEN[2,c] AND LIANG LI[2,d]

[1]*Department of Biostatistics, Columbia University,* [a]*sw3763@cumc.columbia.edu*

[2]*Department of Biostatistics, University of Texas MD Anderson Cancer Center,* [b]*jning@mdanderson.org,* [c]*yshen@mdanderson.org,* [d]*lli15@mdanderson.org*

[3]*Department of Health Services Research, University of Texas MD Anderson Cancer Center ,* [e]*yingxu@mdanderson.org,* [f]*yashih@mdanderson.org*

Insurance claims' data is an increasingly important health policy research resource, given its longitudinal assessment of cancer care clinical outcomes. Population-level information on medical cost trajectory from disease diagnosis to terminal events, such as death, specifically interests policy makers. Estimating the mean cost trajectory has statistical challenges. The shape of the trajectory is usually highly nonlinear with varying durations, depending on the diagnosis-to-death population time distribution. The terminal event may be right censored, resulting in missing subsequent costs. Medical costs often have skewed distributions with zero inflation and heteroscedasticity which may not fit well with the commonly used parametric family of distributions. In this paper we propose a flexible semiparametric model to address challenges without imposing a cost data distributional assumption. The estimation procedure is based on generalized estimating equations with censored covariates. The proposed model adopts a bivariate surface that quantifies the interrelationship between longitudinal medical costs and survival, and results in the nonlinear population mean cost trajectories given survival time. We develop a novel generalized estimating equations algorithm to accommodate covariates subject to right censoring without fully specifying the joint distribution of the cost and survival data. We provide theoretical and simulation-based justification for the proposed approach and apply the methods to estimate prostate cancer patient cost trajectories from the Surveillance, Epidemiology, and End Results (SEER)-Medicare linked database.

REFERENCES

BANG, H. (2005). Medical cost analysis: Application to colorectal cancer data from the SEER Medicare database. *Contemp. Clin. Trials* **26** 586–597.

BANG, H. and TSIATIS, A. A. (2000). Estimating medical costs with censored data. *Biometrika* **87** 329–343. MR1782482 https://doi.org/10.1093/biomet/87.2.329

BAŞER, O., GARDINER, J. C., BRADLEY, C. J., YÜCE, H. and GIVEN, C. (2006). Longitudinal analysis of censored medical cost data. *Health Econ.* **15** 513–525.

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

BROWN, M. L., RILEY, G. F., SCHUSSLER, N. and ETZIONI, R. (2002). Estimating health care costs related to cancer treatment from SEER-Medicare data. *Med. Care* **40** IV104–IV117. https://doi.org/10.1097/00005650-200208001-00014

CAI, T., HYNDMAN, R. J. and WAND, M. P. (2002). Mixed model-based hazard estimation. *J. Comput. Graph. Statist.* **11** 784–798. MR1944263 https://doi.org/10.1198/106186002321018786

CHEN, H. and WANG, Y. (2011). A penalized spline approach to functional mixed effects model analysis. *Biometrics* **67** 861–870. MR2829260 https://doi.org/10.1111/j.1541-0420.2010.01524.x

Key words and phrases. Bivariate penalized spline, generalized estimating equations, joint modeling, medical cost, SEER-Medicare.

CHEN, H., WANG, Y., PAIK, M. C. and CHOI, H. A. (2013b). A marginal approach to reduced-rank penalized spline smoothing with application to multilevel functional data. *J. Amer. Statist. Assoc.* **108** 1216–1229. MR3174699 https://doi.org/10.1080/01621459.2013.826134

CHEN, J., LIU, L., SHIH, Y.-C. T., ZHANG, D. and SEVERINI, T. A. (2016). A flexible model for correlated medical costs, with application to medical expenditure panel survey data. *Stat. Med.* **35** 883–894. MR3457613 https://doi.org/10.1002/sim.6743

CHEN, J., LIU, L., ZHANG, D. and SHIH, Y.-C. T. (2013a). A flexible model for the mean and variance functions, with application to medical cost data. *Stat. Med.* **32** 4306–4318. MR3118356 https://doi.org/10.1002/sim.5838

DE BOOR, C. (1978). *A Practical Guide to Splines*. *Applied Mathematical Sciences* **27**. Springer, New York. MR0507062

EILERS, P. H. C. and MARX, B. D. (2002). Generalized linear additive smooth structures. *J. Comput. Graph. Statist.* **11** 758–783. MR1944262 https://doi.org/10.1198/106186002321018777

EILERS, P. H. and MARX, B. D. (2010). Splines, knots, and penalties. *Wiley Interdiscip. Rev.: Comput. Stat.* **2** 637–653.

FITZMAURICE, G., DAVIDIAN, M., VERBEKE, G. and MOLENBERGHS, G. (2008). *Longitudinal Data Analysis*. CRC Press, Boca Raton, FL.

FU, W. J. (2003). Penalized estimating equations. *Biometrics* **59** 126–132. MR1978479 https://doi.org/10.1111/1541-0420.00015

GOLDFELD, K. and WUJCIAK-JENS, J. (2020). simstudy: Illuminating research methods through data generation. *J. Open Sour. Softw.* **5** 2763.

KAUERMANN, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Comput. Statist. Data Anal.* **49** 169–186. MR2129171 https://doi.org/10.1016/j.csda.2004.05.006

KONG, S., NAN, B., KALBFLEISCH, J. D., SARAN, R. and HIRTH, R. (2018). Conditional modeling of longitudinal data with terminal event. *J. Amer. Statist. Assoc.* **113** 357–368. MR3803470 https://doi.org/10.1080/01621459.2016.1255637

LI, L. and GREENE, T. (2008). Varying coefficients model with measurement error. *Biometrics* **64** 519–526. MR2432422 https://doi.org/10.1111/j.1541-0420.2007.00921.x

LI, L., WU, C.-H., NING, J., HUANG, X., SHIH, Y.-C. T. and SHEN, Y. (2018). Semiparametric estimation of longitudinal medical cost trajectory. *J. Amer. Statist. Assoc.* **113** 582–592. MR3832210 https://doi.org/10.1080/01621459.2017.1361329

LI, Y. and RUPPERT, D. (2008). On the asymptotics of penalized splines. *Biometrika* **95** 415–436. MR2521591 https://doi.org/10.1093/biomet/asn010

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430 https://doi.org/10.1093/biomet/73.1.13

LIANG, K.-Y., ZEGER, S. L. and QAQISH, B. (1992). Multivariate regression analyses for categorical data. *J. Roy. Statist. Soc. Ser. B* **54** 3–24. MR1157713

LIN, D. (2003). Regression analysis of incomplete medical cost data. *Stat. Med.* **22** 1181–1200.

LIN, D. Y. (2000). Linear regression analysis of censored medical costs. *Biostatistics* **1** 35–47. https://doi.org/10.1093/biostatistics/1.1.35

LIN, D., FEUER, E., ETZIONI, R. and WAX, Y. (1997). Estimating medical costs from incomplete follow-up data. *Biometrics* **53** 419–434.

LIU, L. (2009). Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Stat. Med.* **28** 972–986. MR2518360 https://doi.org/10.1002/sim.3497

LIU, L., CONAWAY, M. R., KNAUS, W. A. and BERGIN, J. D. (2008). A random effects four-part model, with application to correlated medical costs. *Comput. Statist. Data Anal.* **52** 4458–4473. MR2432473 https://doi.org/10.1016/j.csda.2008.02.034

LIU, L., HUANG, X. and O'QUIGLEY, J. (2008). Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics* **64** 950–958. MR2526647 https://doi.org/10.1111/j.1541-0420.2007.00954.x

LIU, L., WOLFE, R. A. and KALBFLEISCH, J. D. (2007). A shared random effects model for censored medical costs and mortality. *Stat. Med.* **26** 139–155. MR2312704 https://doi.org/10.1002/sim.2535

MARIOTTO, A. B., ROBIN YABROFF, K., SHAO, Y., FEUER, E. J. and BROWN, M. L. (2011). Projections of the cost of cancer care in the United States: 2010–2020. *J. Natl. Cancer Inst.* **103** 117–128.

SHIH, Y.-C. T., XU, Y., CHIEN, C.-R., KIM, B., SHEN, Y., LI, L. and GEYNISMAN, D. M. (2019). Rising economic burden of renal cell carcinoma among elderly patients in the USA: Part II – An updated analysis of SEER-Medicare data. *PharmacoEcon.* **37** 1495–1507.

SHUN, Z. and MCCULLAGH, P. (1995). Laplace approximation of high-dimensional integrals. *J. Roy. Statist. Soc. Ser. B* **57** 749–760. MR1354079

WANG, S., NING, J., XU, Y., SHIH, Y.-C., SHEN, Y. and LI, L. (2023). Supplement to "An extension of estimating equations to model longitudinal medical cost trajectory with Medicare claims data linked to SEER cancer registry." https://doi.org/10.1214/22-AOAS1659SUPPA, https://doi.org/10.1214/22-AOAS1659SUPPB

WELSH, A. H., LIN, X. and CARROLL, R. J. (2002). Marginal longitudinal nonparametric regression: Locality and efficiency of spline and kernel methods. *J. Amer. Statist. Assoc.* **97** 482–493. MR1941465 https://doi.org/10.1198/016214502760047014

YABROFF, K. R., WARREN, J. L., SCHRAG, D., MARIOTTO, A., MEEKINS, A. TOPOR, M. and BROWN, M. L. (2008). Comparison of approaches for estimating incidence costs of care for colorectal cancer patients. *Med. Care* **47** S56–S63.

ZHAO, H. and TIAN, L. (2001). On estimating medical cost and incremental cost-effectiveness ratios with censored data. *Biometrics* **57** 1002–1008. MR1950417 https://doi.org/10.1111/j.0006-341X.2001.01002.x

ZHAO, H., CHENG, Y. and BANG, H. (2011). Some insight on censored cost estimators. *Stat. Med.* **30** 2381–2388. MR2829139 https://doi.org/10.1002/sim.4295

ZHAO, H., ZUO, C., CHEN, S. and BANG, H. (2012). Nonparametric inference for median costs with censored data. *Biometrics* **68** 717–725. MR3055176 https://doi.org/10.1111/j.1541-0420.2012.01755.x

ZHAO, L. P. and PRENTICE, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77** 642–648. MR1087856 https://doi.org/10.1093/biomet/77.3.642

ZHU, Z., FUNG, W. K. and HE, X. (2008). On the asymptotics of marginal regression splines with longitudinal data. *Biometrika* **95** 907–917. MR2461219 https://doi.org/10.1093/biomet/asn041