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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles

continued

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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles—*Continued from front cover*

THE ANNALS OF APPLIED STATISTICS Vol. 16, No. 2, pp. 627–1252 June 2022

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

The purpose of the Institute is to foster the development and dissemination of the theory and applications of statistics and probability.

IMS OFFICERS

IMS PUBLICATIONS

The Annals of Applied Statistics **[ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 16, Number 2, June 2022. 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
2022, Vol. 16, No. 2, 627–651
https://doi.org/10.1214/21-AOAS1463

MODELLED APPROXIMATIONS TO THE IDEAL FILTER WITH APPLICATION TO GDP AND ITS COMPONENTS[1]

BY THOMAS M. TRIMBUR[a] AND TUCKER S. MCELROY[b]

Center for Statistical Research and Methodology, U.S. Census Bureau, [a]*Thomas.Trimbur@census.gov,* [b]*tucker.s.mcelroy@census.gov*

This paper examines cyclical fluctuations in a comprehensive statistical application, focusing on U.S. macroeconomic indicators related to real gross domestic product (real GDP). While GDP is generally viewed as the most widespread measure of economic activity available, our dataset also encompasses the primary GDP components, such as investment, together with leading (and regularly analyzed) subcomponent series, like residential and inventory investment. Analysis of the cycles in these major sectors provides a more informative perspective on the macroeconomic state and may improve a researcher's ability to understand and forecast cyclical movements and growth in GDP. Adaptive time series modelling is used for each time series to derive the preferred band-pass filter for computing the optimal cycle. This contrasts with the rigid use of the ideal filter, whose gain function is perfectly sharp. Regarding the ideal filter, we provide an improved implementation compared to current practice. Thus, a set of approximating filters is derived that allow for a more attractive gain profile, a better match to the targeted passband, and a direct statistical way to extract signals near the sample endpoints. Our application study demonstrates that the commonly used ideal filter can perform quite poorly on a routine basis and lead to incorrect conclusions about even the most basic questions about empirical cyclical properties. The amplitude of filtered economic activity can have major distortions and become expanded or diminished (depending on the GDP component under consideration), and many essential divergences in path may occur and affect key signals, such as expansion or contraction in growth. Statistical measures of model performance very strongly favor the adaptive parameter approach. Our statistical analysis reveals diverse dynamic behavior among the series; such results may yield worthwhile insights for output sector analysts and, even for those primarily focused on GDP, may lead to possible modelling improvements by using the finer information content in the GDP-component dynamics.

REFERENCES

BASHAR, O., BHATTACHARYA, P. and WOHAR, M. (2017). The cyclicality of fiscal policy: New evidence from unobserved components approach. *Journal of Macroeconomics* **53** 222–234.

BAXTER, M. and KING, R. G. (1999). Measuring business cycles: Approximate band-pass filters for economic time series. *Rev. Econ. Stat.* **81** 575–93.

BELL, W. (1984). Signal extraction for nonstationary time series. *Ann. Statist.* **12** 646–664. MR0740918 https://doi.org/10.1214/aos/1176346512

BULLIGAN, G., BURLON, L., DELLE MONACHE, D. and SILVESTRINI, A. (2019). Real and financial cycles: Estimates using unobserved component models for the Italian economy. *Stat. Methods Appl.* **28** 541–569. MR4009761 https://doi.org/10.1007/s10260-019-00453-1

BURNS, F. and MITCHELL, C. (1946). *Measuring Business Cycles*. National Bureau of Economic Research.

BUSETTI, F. and CAIVANO, M. (2016). The trend-cycle decomposition of output and the Phillips Curve: Bayesian estimates for Italy and the Euro Area. *Empirical Economics* **50** 1565–1587.

CHEN, X. and MILLS, T. C. (2012). Measuring the Euro area output gap using a multivariate unobserved components model containing phase shifts. *Empirical Economics* **43** 671–692.

Key words and phrases. Business cycles, band-pass filter, ideal filter, signal extraction, stochastic cycles, unobserved components.

DONADELLI, M., PARADISO, A. and LIVIERI, G. (2019). Adding cycles into the neoclassical growth model. *Econ. Model.* **78** 162–171.

DOORNIK, J. (2006). *Ox*: *Object-Oriented Matrix Programming Language*. Timberlake Consultants, London.

GONZALEZ, R. B. and MARINHO, L. S. G. (2017). Re-anchoring countercyclical capital buffers: Bayesian estimates and alternatives focusing on credit growth. *Int. J. Forecast.* **33** 1007–1024.

HARVEY, A. C. (1989). *Forecasting*, *Structural Time Series Models and the Kalman Filter*. Cambridge Univ. Press, Cambridge.

HARVEY, A. C. and TRIMBUR, T. (2003). General model-based filters for extracting cycles and trends in economic time series. *Rev. Econ. Stat.* **85** 244–55.

HARVEY, A. C., TRIMBUR, T. M. and VAN DIJK, H. K. (2007). Trends and cycles in economic time series: A Bayesian approach. *J. Econometrics* **140** 618–649. MR2408920 https://doi.org/10.1016/j.jeconom.2006.07.006

RÜNSTLER, G. and VLEKKE, M. (2018). Business, housing, and credit cycles. *J. Appl. Econometrics* **33** 212–226. MR3782759 https://doi.org/10.1002/jae.2604

TAWADROS, G. (2013). The cyclicality of the demand for crude oil: Evidence from the OECD. *Journal of Economic Studies* **40** 704–719.

TRIMBUR, T. M. and MCELROY, T. S. (2022). Supplement to "Modelled approximations to the ideal filter with application to GDP and its components." https://doi.org/10.1214/21-AOAS1463SUPPA, https://doi.org/10.1214/21-AOAS1463SUPPB

WOLFRAM, S. (1996). *The Mathematica Book*, 3rd ed. Wolfram Media, Champaign, IL. MR1404696

ZIZZA, R. (2006). A measure of output gap for Italy through structural time series models. *J. Appl. Stat.* **33** 481–495. MR2227407 https://doi.org/10.1080/02664760500448875

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 652–679
https://doi.org/10.1214/21-AOAS1471

IMPROVING EXOPLANET DETECTION POWER: MULTIVARIATE GAUSSIAN PROCESS MODELS FOR STELLAR ACTIVITY

BY DAVID E. JONES[1,a], DAVID C. STENNING[2,b], ERIC B. FORD[3,4,5,6,c], ROBERT L. WOLPERT[7,e], THOMAS J. LOREDO[8,f], CHRISTIAN GILBERTSON[3,d] AND XAVIER DUMUSQUE[9,g]

[1]*Department of Statistics, Texas A&M University,* [a]*david.jones@tamu.edu*
[2]*Department of Statistics and Actuarial Science, Simon Fraser University,* [b]*dstennin@sfu.ca*
[3]*Department of Astronomy and Astrophysics, Pennsylvania State University,* [c]*ebf11@psu.edu,* [d]*cjg66@psu.edu*
[4]*Center for Exoplanets & Habitable Worlds, Pennsylvania State University*
[5]*Center for Astrostatistics, Pennsylvania State University*
[6]*Institute for Computational & Data Sciences, Pennsylvania State University*
[7]*Department of Statistical Science, Duke University,* [e]*rlw@duke.edu*
[8]*Cornell Center for Astrophysics and Planetary Science, Cornell University,* [f]*loredo@astro.cornell.edu*
[9]*Observatoire Astronomique de l'Université de Genève,* [g]*Xavier.Dumusque@unige.ch*

The radial velocity method is one of the most successful techniques for detecting exoplanets. It works by detecting the velocity of a host star, induced by the gravitational effect of an orbiting planet, specifically, the velocity along our line of sight which is called the *radial velocity* of the star. Low-mass planets typically cause their host star to move with radial velocities of 1 m/s or less. By analyzing a time series of stellar spectra from a host star, modern astronomical instruments can, in theory, detect such planets. However, in practice, intrinsic stellar variability (e.g., star spots, convective motion, pulsations) affects the spectra and often mimics a radial velocity signal. This signal contamination makes it difficult to reliably detect low-mass planets. A principled approach to recovering planet radial velocity signals in the presence of stellar activity was proposed by Rajpaul et al. (*Mon. Not. R. Astron. Soc.* **452** (2015) 2269–2291). It uses a multivariate Gaussian process model to jointly capture time series of the apparent radial velocity and multiple indicators of stellar activity. We build on this work in two ways: (i) we propose using dimension reduction techniques to construct new high-information stellar activity indicators; and (ii) we extend the Rajpaul et al. (*Mon. Not. R. Astron. Soc.* **452** (2015) 2269–2291) model to a larger class of models and use a power-based model comparison procedure to select the best model. Despite significant interest in exoplanets, previous efforts have not performed large-scale stellar activity model selection or attempted to evaluate models based on planet detection power. In the case of main sequence G2V stars, we find that our method substantially improves planet detection power, compared to previous state-of-the-art approaches.

REFERENCES

ADLER, R. J. (2010). *The Geometry of Random Fields*. *Classics in Applied Mathematics* **62**. SIAM, Philadelphia, PA. MR3396215 https://doi.org/10.1137/1.9780898718980.ch1

AIGRAIN, S., PONT, F. and ZUCKER, S. (2012). A simple method to estimate radial velocity variations due to stellar activity using photometry. *Mon. Not. R. Astron. Soc.* **419** 3147–3158.

ÁLVAREZ, M. A. and LAWRENCE, N. D. (2011). Computationally efficient convolved multiple output Gaussian processes. *J. Mach. Learn. Res.* **12** 1459–1500. MR2813145

ARENTOFT, T., KJELDSEN, H., BEDDING, T. R., BAZOT, M., CHRISTENSEN-DALSGAARD, J., DALL, T. H., KAROFF, C., CARRIER, F., EGGENBERGER, P. et al. (2008). A multisite campaign to measure solar-like oscillations in procyon. I. Observations, data reduction, and slow variations. *Astrophys. J.* **687** 1180.

Key words and phrases. Exoplanets, radial velocity method, stellar activity, Gaussian process, time series, dimension reduction, model selection.

BARANNE, A., QUELOZ, D., MAYOR, M., ADRIANZYK, G., KNISPEL, G., KOHLER, D., LACROIX, D., MEUNIER, J.-P., RIMBAUD, G. et al. (1996). ELODIE: A spectrograph for accurate radial velocity measurements. *Astron. Astrophys. Suppl. Ser.* **119** 373–390.

BAUMANN, I. and SOLANKI, S. K. (2005). On the size distribution of sunspot groups in the Greenwich sunspot record 1874–1976. *Astron. Astrophys.* **443** 1061–1066. https://doi.org/10.1051/0004-6361:20053415

BOISSE, I., BONFILS, X. and SANTOS, N. (2012). SOAP: A tool for the fast computation of photometry and radial velocity induced by stellar spots. *Astron. Astrophys.* **545** A109.

BORGNIET, S., MEUNIER, N. and LAGRANGE, A.-M. (2015). Using the Sun to estimate Earth-like planets detection capabilities—V. Parameterizing the impact of solar activity components on radial velocities. *Astron. Astrophys.* **581** A133.

BUTLER, R. P., MARCY, G. W., WILLIAMS, E., MCCARTHY, C., DOSANJH, P. and VOGT, S. S. (1996). Attaining Doppler precision of 3 M s-1. *Publ. Astron. Soc. Pac.* **108** 500.

BUTLER, R. P., VOGT, S. S., LAUGHLIN, G., BURT, J. A., RIVERA, E. J., TUOMI, M., TESKE, J., ARRIAGADA, P., DIAZ, M. et al. (2017). The LCES HIRES/Keck precision radial velocity exoplanet survey. *Astron. J.* **153** 208.

COIFMAN, R. R. and LAFON, S. (2006). Diffusion maps. *Appl. Comput. Harmon. Anal.* **21** 5–30. MR2238665 https://doi.org/10.1016/j.acha.2006.04.006

COIFMAN, R. R., LAFON, S., LEE, A. B., MAGGIONI, M., NADLER, B., WARNER, F. and ZUCKER, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. USA* **102** 7426–7431.

DAMASSO, M., PINAMONTI, M., SCANDARIATO, G. and SOZZETTI, A. (2019). Biases in retrieving planetary signals in the presence of quasi-periodic stellar activity. *Mon. Not. R. Astron. Soc.* **489** 2555–2571.

DANBY, J. M. A. (1988). *Fundamentals of Celestial Mechanics*, 2nd ed. Willmann-Bell, Inc., Richmond, VA. MR0981360

DAVIS, A. B., CISEWSKI, J., DUMUSQUE, X., FISCHER, D. A. and FORD, E. B. (2017). Insights on the spectral signatures of stellar activity and planets from PCA. *Astrophys. J.* **846** 59.

DEL MORO, D. (2004). Solar granulation properties derived from three different time series. *Astron. Astrophys.* **428** 1007–1015.

DEL MORO, D., BERRILLI, F., DUVALL, T. and KOSOVICHEV, A. (2004). Dynamics and structure of supergranulation. *Sol. Phys.* **221** 23–32.

DUMUSQUE, X. (2016). Radial velocity fitting challenge—I. Simulating the data set including realistic stellar radial-velocity signals. *Astron. Astrophys.* **593** A5.

DUMUSQUE, X., BOISSE, I. and SANTOS, N. (2014). SOAP 2.0: A tool to estimate the photometric and radial velocity variations induced by stellar spots and plages. *Astrophys. J.* **796** 132.

DUMUSQUE, X., PEPE, F., LOVIS, C., SÉGRANSAN, D., SAHLMANN, J., BENZ, W., BOUCHY, F., MAYOR, M., QUELOZ, D. et al. (2012). An Earth-mass planet orbiting α Centauri B. *Nature* **491** 207–211.

DUMUSQUE, X., BORSA, F., DAMASSO, M., DIAZ, R. F., GREGORY, P., HARA, N., HATZES, A., RAJPAUL, V., TUOMI, M. et al. (2017). Radial-velocity fitting challenge—II. First results of the analysis of the data set. *Astron. Astrophys.* **598** A133.

DUPUY, D., HELBERT, C., FRANCO, J. et al. (2015). DiceDesign and DiceEval: Two R packages for design and analysis of computer experiments. *J. Stat. Softw.* **65** 1–38.

FISCHER, D. A., MARCY, G. W. and SPRONCK, J. F. (2013). The twenty-five year Lick planet search. *Astrophys. J., Suppl. Ser.* **210** 5.

FISCHER, D. A., ANGLADA-ESCUDE, G., ARRIAGADA, P., BALUEV, R. V., BEAN, J. L., BOUCHY, F., BUCHHAVE, L. A., CARROLL, T., CHAKRABORTY, A. et al. (2016). State of the field: Extreme precision radial velocities. *Publ. Astron. Soc. Pac.* **128** 066001.

FORD, E. B. (2006). Improving the efficiency of Markov chain Monte Carlo for analyzing the orbits of extrasolar planets. *Astrophys. J.* **642** 505–522. https://doi.org/10.1086/500802

HAYWOOD, R. D., COLLIER CAMERON, A., UNRUH, Y., LOVIS, C., LANZA, A., LLAMA, J., DELEUIL, M., FARES, R., GILLON, M. et al. (2016). The Sun as a planet-host star: Proxies from SDO images for HARPS radial-velocity variations. *Mon. Not. R. Astron. Soc.* **457** 3637–3651.

JENKINS, J. L. (2013). *Observing the Sun*: *A Pocket Field Guide*. Springer, New York.

JONES, D., STENNING, D., EB, F., WOLPERT, R., LOREDO, T., GILBERTSON, C. and DUMUSQUE, X. (2022). Supplement to "Improving exoplanet detection power: Multivariate Gaussian process models for stellar activity." https://doi.org/10.1214/21-AOAS1471SUPP

JOURNEL, A. G. and HUIJBREGTS, C. J. (1978). *Mining Geostatistics*. Academic Press, San Diego.

LOREDO, T. J., BERGER, J. O., CHERNOFF, D. F., CLYDE, M. A. and LIU, B. (2012). Bayesian methods for analysis and adaptive scheduling of exoplanet observations. *Stat. Methodol.* **9** 101–114. MR2863601 https://doi.org/10.1016/j.stamet.2011.07.005

MANDAL, S., KARAK, B. B. and BANERJEE, D. (2017). Latitude distribution of sunspots: Analysis using sunspot data and a dynamo model. *Astrophys. J.* **851** 70. https://doi.org/10.3847/1538-4357/aa97dc

MAYOR, M., PEPE, F., QUELOZ, D., BOUCHY, F., RUPPRECHT, G., LO CURTO, G., AVILA, G., BENZ, W., BERTAUX, J. L. et al. (2003). Setting new standards with HARPS. *Messenger* **114** 20–24.

MAYOR, M., MARMIER, M., LOVIS, C., UDRY, S., SÉGRANSAN, D., PEPE, F., BENZ, W., BERTAUX, J.-L., BOUCHY, F. et al. (2011). The HARPS search for southern extra-solar planets XXXIV. Occurrence, mass distribution and orbital properties of super-Earths and Neptune-mass planets. Preprint. Available at arXiv:1109.2497.

MORRIS, M. D. and MITCHELL, T. J. (1995). Exploratory designs for computational experiments. *J. Statist. Plann. Inference* **43** 381–402. https://doi.org/10.1016/0378-3758(94)00035-T

OSBORNE, M. A., ROBERTS, S. J., ROGERS, A., RAMCHURN, S. D. and JENNINGS, N. R. (2008). Towards real-time information processing of sensor network data using computationally efficient multi-output Gaussian processes. In *Proceedings of the 7th International Conference on Information Processing in Sensor Networks* 109–120. IEEE Computer Society, Los Alamitos.

PEPE, F., LOVIS, C., SEGRANSAN, D., BENZ, W., BOUCHY, F., DUMUSQUE, X., MAYOR, M., QUELOZ, D., SANTOS, N. et al. (2011). The HARPS search for Earth-like planets in the habitable zone—I. Very low-mass planets around HD 20794, HD 85512, and HD 192310. *Astron. Astrophys.* **534** A58.

RAJPAUL, V., AIGRAIN, S. and ROBERTS, S. (2015). Ghost in the time series: No planet for Alpha Cen B. *Mon. Not. R. Astron. Soc. Lett.* **456** L6–L10.

RAJPAUL, V., AIGRAIN, S., OSBORNE, M. A., REECE, S. and ROBERTS, S. (2015). A Gaussian process framework for modelling stellar activity signals in radial velocity data. *Mon. Not. R. Astron. Soc.* **452** 2269–2291.

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

WALLACE, L., HINKLE, K. and LIVINGSTON, W. (1998). An atlas of the spectrum of the solar photosphere from 13,500 to 28,000 cm^{-1} (3570 to 7405 Å). *Astrophys. Space Sci. Proc.*

WALLACE, L., HINKLE, K. and LIVINGSTON, W. (2005). An atlas of sunspot umbral spectra in the visible from 15,000 to 25,500 cm^{-1} (3920 to 6664 Å).

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 680–705
https://doi.org/10.1214/21-AOAS1485

A BAYESIAN MODEL OF DOSE-RESPONSE FOR CANCER DRUG STUDIES

BY WESLEY TANSEY[1,a], CHRISTOPHER TOSH[2,b] AND DAVID M. BLEI[2,c]

[1]*Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center,* [a]*tanseyw@mskcc.org*
[2]*Data Science Institute, Columbia University,* [b]*ct2915@columbia.edu,* [c]*david.blei@columbia.edu*

Exploratory cancer drug studies test multiple tumor cell lines against multiple candidate drugs. The goal in each paired (cell line, drug) experiment is to map out the dose-response curve of the cell line as the dose level of the drug increases. We propose Bayesian tensor filtering (BTF), a hierarchical Bayesian model for dose-response modeling in multisample, multitreatment cancer drug studies. BTF uses low-dimensional embeddings to share statistical strength between similar drugs and similar cell lines. Structured shrinkage priors in BTF encourage smoothness in the dose-response curves while remaining adaptive to sharp jumps when the data call for it. We focus on a pair of cancer drug studies exhibiting a particular pathology in their experimental design, leading us to a nonconjugate monotone mixture-of-gammas likelihood. To perform posterior inference, we develop a variant of the elliptical slice sampling algorithm for sampling from linearly-constrained multivariate normal priors with nonconjugate likelihoods. In benchmarks, BTF outperforms state-of-the-art methods for covariance regression and dynamic Poisson matrix factorization. On the two cancer drug studies, BTF outperforms the current standard approach in biology and reveals potential new biomarkers of drug sensitivity in cancer. Code is available at https://github.com/tansey/functionalmf.

REFERENCES

ABBAS-AGHABABAZADEH, F., LU, P. and FRIDLEY, B. L. (2019). Nonlinear mixed-effects models for modeling *in vitro* drug response data to determine problematic cancer cell lines. *Sci. Rep.* **9** 1–9.

AN, Z., AKSOY, O., ZHENG, T., FAN, Q.-W. and WEISS, W. A. (2018). Epidermal growth factor receptor and EGFRvIII in glioblastoma: Signaling pathways and targeted therapies. *Oncogene* **37** 1561–1575.

BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Anal.* **12** 1105–1131. MR3724980 https://doi.org/10.1214/16-BA1028

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

BORNKAMP, B. and ICKSTADT, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis. *Biometrics* **65** 198–205. MR2665861 https://doi.org/10.1111/j.1541-0420.2008.01060.x

CAI, B. and DUNSON, D. B. (2007). Bayesian multivariate isotonic regression splines: Applications to carcinogenicity studies. *J. Amer. Statist. Assoc.* **102** 1158–1171. MR2412540 https://doi.org/10.1198/016214506000000942

CANONICI, A., GIJSEN, M., MULLOOLY, M., BENNETT, R., BOUGUERN, N., PEDERSEN, K., O'BRIEN, N. A., ROXANIS, I., LI, J.-L. et al. (2013). Neratinib overcomes trastuzumab resistance in HER2 amplified breast cancer. *Oncotarget* **4** 1592.

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

CELEUX, G., FORBES, F., ROBERT, C. P. and TITTERINGTON, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Anal.* **1** 651–673. MR2282197 https://doi.org/10.1214/06-BA122

DROST, J. and CLEVERS, H. (2018). Organoids in cancer research. *Nat. Rev. Cancer* **18** 407–418. https://doi.org/10.1038/s41568-018-0007-6

FAGAN, F., BHANDARI, J. and CUNNINGHAM, J. (2016). Elliptical slice sampling with expectation propagation. In *Uncertainty in Artificial Intelligence*.

Key words and phrases. Dose-response, matrix factorization, trend filtering, slice sampling, constrained inference.

FAULKNER, J. R. and MININ, V. N. (2018). Locally adaptive smoothing with Markov random fields and shrinkage priors. *Bayesian Anal*. **13** 225–252. MR3737950 https://doi.org/10.1214/17-BA1050

FOX, E. B. and DUNSON, D. B. (2015). Bayesian nonparametric covariance regression. *J. Mach. Learn. Res*. **16** 2501–2542. MR3450515

FRIDLEY, B. L., JENKINS, G., SCHAID, D. J. and WANG, L. (2009). A Bayesian hierarchical nonlinear model for assessing the association between genetic variation and drug cytotoxicity. *Stat. Med*. **28** 2709–2722. MR2751045 https://doi.org/10.1002/sim.3649

GARNETT, M. J., EDELMAN, E. J., HEIDORN, S. J., GREENMAN, C. D., DASTUR, A., LAU, K. W., GRENINGER, P., THOMPSON, I. R., LUO, X. et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483** 570.

GAUVIN, L., PANISSON, A. and CATTUTO, C. (2014). Detecting the community structure and activity patterns of temporal networks: A non-negative tensor factorization approach. *PLoS ONE* **9** e86028. https://doi.org/10.1371/journal.pone.0086028

GHANDI, M., HUANG, F. W., JANÉ-VALBUENA, J., KRYUKOV, G. V., LO, C. C., MCDONALD, E. R., BARRETINA, J., GELFAND, E. T., BIELSKI, C. M. et al. (2019). Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569** 503–508.

GHEBRETINSAE, A. H., FAES, C., MOLENBERGHS, G., DE BOECK, M. and GEYS, H. (2013). A Bayesian, generalized frailty model for comet assays. *J. Biopharm. Statist*. **23** 618–636. MR3049131 https://doi.org/10.1080/10543406.2012.756499

GUO, G., GONG, K., WOHLFELD, B., HATANPAA, K. J., ZHAO, D. and HABIB, A. A. (2015). Ligand-independent EGFR signaling. *Cancer Res*. **75** 3436–3441.

HAHN, P. R., HE, J. and LOPES, H. (2018). Bayesian factor model shrinkage for linear IV regression with many instruments. *J. Bus. Econom. Statist*. **36** 278–287. MR3790214 https://doi.org/10.1080/07350015.2016.1172968

HEAUKULANI, C. and VAN DER WILK, M. (2019). Scalable Bayesian dynamic covariance modeling with variational Wishart and inverse Wishart processes. In *Advances in Neural Information Processing Systems* 4582–4592.

HUANG, L., WU, S. and XING, D. (2011). High fluence low-power laser irradiation induces apoptosis via inactivation of akt/GSK3β signaling pathway. *J. Cell. Physiol*. **226** 588–601.

JOHNSON, W. E., LI, C. and RABINOVIC, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8** 118–127.

KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. MR2505584 https://doi.org/10.1137/070690274

KOLCH, W., HALASZ, M., GRANOVSKAYA, M. and KHOLODENKO, B. N. (2015). The dynamic control of signal transduction networks in cancer cells. *Nat. Rev. Cancer* **15** 515–527.

KOUL, D. (2008). PTEN signaling pathways in glioblastoma. *Cancer Biol. Ther.* **7** 1321–1325.

KOWAL, D. R., MATTESON, D. S. and RUPPERT, D. (2019). Dynamic shrinkage processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **81** 781–804. MR3997101 https://doi.org/10.1111/rssb.12325

KUNIHAMA, T., HALPERN, C. T. and HERRING, A. H. (2019). Non-parametric Bayes models for mixed scale longitudinal surveys. *J. R. Stat. Soc. Ser. C. Appl. Stat*. **68** 1091–1109. MR4002385 https://doi.org/10.1111/rssc.12348

KYUNG, M., GILL, J., GHOSH, M. and CASELLA, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal*. **5** 369–411. MR2719657 https://doi.org/10.1214/10-BA607

LACHMANN, A., GIORGI, F. M., ALVAREZ, M. J. and CALIFANO, A. (2016). Detection and removal of spatial bias in multiwell assays. *Bioinformatics* **32** 1959–1965.

LEE, J.-K., LIU, Z., SA, J. K., SHIN, S., WANG, J., BORDYUH, M., CHO, H. J., ELLIOTT, O., CHU, T. et al. (2018). Pharmacogenomic landscape of patient-derived tumor cells informs precision oncology therapy. *Nat. Genet*. **50** 1399–1411.

LEEK, J. T., SCHARPF, R. B., BRAVO, H. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GEMAN, D., BAGGERLY, K. and IRIZARRY, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet*. **11** 733–739.

LI, L., PLUTA, D., SHAHBABA, B., FORTIN, N., OMBAO, H. and BALDI, P. (2019). Modeling dynamic functional connectivity with latent factor Gaussian processes. In *Advances in Neural Information Processing Systems* 8263–8273.

LIN, L. and DUNSON, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika* **101** 303–317. MR3215349 https://doi.org/10.1093/biomet/ast063

LINDGREN, F. and RUE, H. (2008). On the second-order random walk model for irregular locations. *Scand. J. Stat*. **35** 691–700. MR2468870 https://doi.org/10.1111/j.1467-9469.2008.00610.x

LOW-KAM, C., TELESCA, D., JI, Z., ZHANG, H., XIA, T., ZINK, J. I. and NEL, A. E. (2015). A Bayesian regression tree approach to identify the effect of nanoparticles' properties on toxicity profiles. *Ann. Appl. Stat.* **9** 383–401. MR3341120 https://doi.org/10.1214/14-AOAS797

MAZOURE, B., NADON, R. and MAKARENKOV, V. (2017). Identification and correction of spatial bias are essential for obtaining quality data in high-throughput screening technologies. *Sci. Rep.* **7** 11921. https://doi.org/10.1038/s41598-017-11940-4

MINKA, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence* 362–369.

MORAN, K. R., DUNSON, D., WHEELER, M. W. and HERRING, A. H. (2019). Bayesian joint modeling of chemical structure and dose response curves. arXiv preprint arXiv:1912.12228.

MURRAY, I., ADAMS, R. and MACKAY, D. (2010). Elliptical slice sampling. In *Artificial Intelligence and Statistics*.

NAGATA, Y., LAN, K.-H., ZHOU, X., TAN, M., ESTEVA, F. J., SAHIN, A. A., KLOS, K. S., LI, P., MONIA, B. P. et al. (2004). PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients. *Cancer Cell* **6** 117–127.

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

NEELON, B. and DUNSON, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics* **60** 398–406. MR2066274 https://doi.org/10.1111/j.0006-341X.2004.00184.x

PATEL, T., TELESCA, D., GEORGE, S. and NEL, A. E. (2012). Toxicity profiling of engineered nanomaterials via multivariate dose-response surface modeling. *Ann. Appl. Stat.* **6** 1707–1729. MR3058681 https://doi.org/10.1214/12-AOAS563

PERRON, F. and MENGERSEN, K. (2001). Bayesian nonparametric modeling using mixtures of triangular distributions. *Biometrics* **57** 518–528. MR1855686 https://doi.org/10.1111/j.0006-341X.2001.00518.x

PIEGORSCH, W. W., XIONG, H., BHATTACHARYA, R. N. and LIN, L. (2012). Nonparametric estimation of benchmark doses in environmental risk assessment. *Environmetrics* **23** 717–728. MR3019063 https://doi.org/10.1002/env.2175

POLSON, N. G. and SCOTT, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Stat.* **9** 501–538. MR3204017 https://doi.org/10.1093/acprof:oso/9780199694587.003.0017

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

SCHEIN, A., WALLACH, H. and ZHOU, M. (2016). Poisson–Gamma dynamical systems. In *Advances in Neural Information Processing Systems*.

SHIVELY, T. S., SAGER, T. W. and WALKER, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 159–175. MR2655528 https://doi.org/10.1111/j.1467-9868.2008.00677.x

SHOEMAKER, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6** 813–823.

SPIEGEL, S., CLAUSEN, J., ALBAYRAK, S. and KUNEGIS, J. (2011). Link prediction on evolving data using tensor factorization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

TAKEUCHI, K., KASHIMA, H. and UEDA, N. (2017). Autoregressive tensor factorization for spatio-temporal predictions. In *International Conference on Data Mining*.

TANSEY, W., TOSH, C. and BLEI, D. M. (2022). Supplement to "A Bayesian model of dose-response for cancer drug studies." https://doi.org/10.1214/21-AOAS1485SUPPA, https://doi.org/10.1214/21-AOAS1485SUPPB

TANSEY, W., LI, K., ZHANG, H., LINDERMAN, S. W., RABADAN, R., BLEI, D. M. and WIGGINS, C. H. (2021). Dose–response modeling in high-throughput cancer drug screenings: An end-to-end approach. *Biostatistics*.

TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323. MR3189487 https://doi.org/10.1214/13-AOS1189

VAN DER PAS, S. L., KLEIJN, B. J. K. and VAN DER VAART, A. W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electron. J. Stat.* **8** 2585–2618. MR3285877 https://doi.org/10.1214/14-EJS962

VIS, D. J., BOMBARDELLI, L., LIGHTFOOT, H., IORIO, F., GARNETT, M. J. and WESSELS, L. F. (2016). Multilevel models improve precision and speed of IC50 estimates. *Pharmacogenomics J.* **17** 691–700. https://doi.org/10.2217/pgs.16.15

WANG, Y.-X., SMOLA, A. and TIBSHIRANI, R. (2014). The falling factorial basis and its statistical applications. In *International Conference on Machine Learning* 730–738.

WEINSTEIN, J. N., COLLISSON, E. A., MILLS, G. B., SHAW, K. R. M., OZENBERGER, B. A., ELLROTT, K., SHMULEVICH, I., SANDER, C., STUART, J. M. et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45** 1113.

WEST, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) 733–742. Oxford Univ. Press, New York. MR2003537

WHEELER, M. W. (2019). Bayesian additive adaptive basis tensor product models for modeling high dimensional surfaces: An application to high-throughput toxicity testing. *Biometrics* **75** 193–201. MR3953720

WILSON, A., REIF, D. M. and REICH, B. J. (2014). Hierarchical dose-response modeling for high-throughput toxicity screening of environmental chemicals. *Biometrics* **70** 237–246. MR3251684 https://doi.org/10.1111/biom.12114

XIONG, L., CHEN, X., HUANG, T.-K., SCHNEIDER, J. and CARBONELL, J. G. (2010). Temporal collaborative filtering with Bayesian probabilistic tensor factorization. In *International Conference on Data Mining*.

ZHANG, A. and PAISLEY, J. (2018). Deep Bayesian nonparametric tracking. In *International Conference on Machine Learning* 5828–5836.

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 706–720
https://doi.org/10.1214/21-AOAS1502

THE ASSESSMENT OF REPLICATION SUCCESS BASED ON RELATIVE EFFECT SIZE

BY LEONHARD HELD[a], CHARLOTTE MICHELOUD[b] AND SAMUEL PAWEL[c]

Epidemiology, Biostatistics and Prevention Institute, Center for Reproducible Science, University of Zurich,
[a]*leonhard.held@uzh.ch,* [b]*charlotte.micheloud@uzh.ch,* [c]*samuel.pawel@uzh.ch*

Replication studies are increasingly conducted in order to confirm original findings. However, there is no established standard how to assess replication success, and, in practice, many different approaches are used. The purpose of this paper is to refine and extend a recently proposed reverse-Bayes approach for the analysis of replication studies. We show how this method is directly related to the relative effect size, the ratio of the replication to the original effect estimate. This perspective leads to a new proposal to recalibrate the assessment of replication success, the golden level. The recalibration ensures that, for borderline significant original studies, replication success can only be achieved if the replication effect estimate is larger than the original one. Conditional power for replication success can then take any desired value if the original study is significant and the replication sample size is large enough. Compared to the standard approach to require statistical significance of both the original and replication study, replication success at the golden level offers uniform gains in project power and controls the type-I error rate if the replication sample size is not smaller than the original one. An application to data from four large replication projects shows that the new approach leads to more appropriate inferences, as it penalizes shrinkage of the replication estimate, compared to the original one, while ensuring that both effect estimates are sufficiently convincing on their own.

REFERENCES

ANDERSON, S. F. and MAXWELL, S. E. (2017). Addressing the "Replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivar. Behav. Res.* **52** 305–324. https://doi.org/10.1080/00273171.2017.1289361

BALAFOUTAS, L. and SUTTER, M. (2012). Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science* **335** 579–582. https://doi.org/10.1126/science.1211180.

BEGLEY, C. G. and IOANNIDIS, J. P. A. (2015). Reproducibility in science. *Circ. Res.* **116** 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819.

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* **143** 383–430. MR0603745 https://doi.org/10.2307/2982063

CAMERER, C. F., DREBER, A., FORSELL, E., HO, T. H., HUBER, J., JOHANNESSON, M., KIRCHLER, M., ALMENBERG, J., ALTMEJD, A. et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science* **351** 1433–1436. https://doi.org/10.1126/science.aaf0918.

CAMERER, C. F., DREBER, A., HOLZMEISTER, F., HO, T.-H., HUBER, J., JOHANNESSON, M., KIRCHLER, M., NAVE, G., NOSEK, B. A. et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat. Hum. Behav.* **2** 637–644. https://doi.org/10.1038/s41562-018-0399-z.

COVA, F., STRICKLAND, B., ABATISTA, A., ALLARD, A., ANDOW, J., ATTIE, M., BEEBE, J., BERNIŪNAS, R., BOUDESSEUL, J. et al. (2018). Estimating the reproducibility of experimental philosophy. *Rev. Philos. Psychol*. https://doi.org/10.1007/s13164-018-0400-9.

DAWID, A. P. (1982). The well-calibrated Bayesian. *J. Amer. Statist. Assoc.* **77** 605–613. MR0675887 https://doi.org/10.1080/01621459.1982.10477856

Key words and phrases. Power, replication studies, sceptical p-value, shrinkage, two-trials rule, type-I error rate.

EBERSOLE, C. R., ATHERTON, O. E., BELANGER, A. L., SKULBORSTAD, H. M., ALLEN, J. M., BANKS, J. B., BARANSKI, E., BERNSTEIN, M. J., BONFIGLIO, D. B. V. et al. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67** 68–82. https://doi.org/10.1016/j.jesp.2015.10.012.

ERRINGTON, T. M., IORNS, E., GUNN, W., TAN, F. E., LOMAX, J. and NOSEK, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *eLife* **3**. https://doi.org/10.7554/eLife.04333

FDA (1998). Providing clinical evidence of effectiveness for human drug and biological products.

FISHER, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron* **1** 3–32. https://doi.org/10.2307/2331802.

GRIEVE, A. P. (2016). Idle thoughts of a 'well-calibrated' Bayesian in clinical drug development. *Pharm. Stat.* **15** 96–108. https://doi.org/10.1002/pst.1736

HELD, L. (2020a). A new standard for the analysis and design of replication studies (with discussion). *J. Roy. Statist. Soc. Ser. A* **183** 431–469. MR4052785 https://doi.org/10.1111/rssa.12493

HELD, L. (2020b). The harmonic mean χ^2-test to substantiate scientific findings. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 697–708. MR4098969 https://doi.org/10.1111/rssc.12410

IOANNIDIS, J. P. A. (2005). Why most published research findings are false. *PLoS Med.* **2** e124. https://doi.org/10.1371/journal.pmed.0020124

JOHNSON, V. E., PAYNE, R. D., WANG, T., ASHER, A. and MANDAL, S. (2017). On the reproducibility of psychological science. *J. Amer. Statist. Assoc.* **112** 1–10. MR3646548 https://doi.org/10.1080/01621459.2016.1240079

KAY, R. (2015). *Statistical Thinking for Non-statisticians in Drug Regulation*, 2nd ed. Wiley, Chichester, UK. https://doi.org/10.1002/9781118451885.

KLEIN, R. A., RATLIFF, K. A., VIANELLO, M., ADAMS, R. B., BAHNÍK, Š., BERNSTEIN, M. J., BOCIAN, K., BRANDT, M. J., BROOKS, B. et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Soc. Psychol.* **45** 142–152. https://doi.org/10.1027/1864-9335/a000178.

KLEIN, R. A., VIANELLO, M., HASSELMAN, F., ADAMS, B. G., ADAMS, R. B. JR., ALPER, S., AVEYARD, M., AXT, J. R., BABALOLA, M. T. et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* **1** 443–490. https://doi.org/10.1177/2515245918810225.

LY, A. and WAGENMAKERS, E. J. (2020). Discussion of "A new standard for the analysis and design of replication studies" by Leonhard Held. *J. Roy. Statist. Soc. Ser. A* **183** 460–461. https://doi.org/10.1111/rssa.12544.

MACA, J., GALLO, P., BRANSON, M. and MAURER, W. (2002). Reconsidering some aspects of the two-trials paradigm. *J. Biopharm. Statist.* **12** 107–119. https://doi.org/10.1081/bip-120006450.

MATTHEWS, R. A. J. (2001a). Methods for assessing the credibility of clinical trial outcomes. *Drug Inf. J.* **35** 1469–1478. https://doi.org/10.1177/009286150103500442.

MATTHEWS, R. A. J. (2001b). Why should clinicians care about Bayesian methods? *J. Statist. Plann. Inference* **94** 43–58. MR1820171 https://doi.org/10.1016/S0378-3758(00)00232-9

MATTHEWS, J. N. S. (2006). *Introduction to Randomized Controlled Clinical Trials*, 2nd ed. *Texts in Statistical Science Series*. CRC Press/CRC, Boca Raton, FL. MR2261274 https://doi.org/10.1201/9781420011302

MICHELOUD, C. and HELD, L. (2021). Power calculations for replication studies. *Statist. Sci.* To appear.

MURADCHANIAN, J., HOEKSTRA, R., KIERS, H. and VAN RAVENZWAAIJ, D. (2021). How best to quantify replication success? A simulation study on the comparison of replication success metrics. *R. Soc. Open Sci.* **8** 201697. https://doi.org/10.1098/rsos.201697.

NICHOLS, S. (2006). Folk intuitions on free will. *J. Cogn. Cult.* **6** 57–86. https://doi.org/10.1163/156853706776931385.

OBERAUER, K. (2008). How to say no: Single- and dual-process theories of short-term recognition tested on negative probes. *J. Exp. Psychol. Learn. Mem. Cogn.* **34** 439–459. https://doi.org/10.1037/0278-7393.34.3.439.

OPEN SCIENCE COLLABORATION (2015). Estimating the reproducibility of psychological science. *Science* **349** aac4716. https://doi.org/10.1126/science.aac4716

PAWEL, S. and HELD, L. (2020). Probabilistic forecasting of replication studies. *PLoS ONE* **15** e0231416. https://doi.org/10.1371/journal.pone.0231416

PAYNE, B. K., BURKLEY, M. A. and STOKES, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *J. Pers. Soc. Psychol.* **94** 16–31. https://doi.org/10.1037/0022-3514.94.1.16

PYC, M. A. and RAWSON, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science* **330** 335. https://doi.org/10.1126/science.1191465

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681 https://doi.org/10.1214/aos/1176346785

SCHMIDT, J. R. and BESNER, D. (2008). The Stroop effect: Why proportion congruent has nothing to do with congruency and everything to do with contingency. *J. Exp. Psychol. Learn. Mem. Cogn.* **34** 514–523. https://doi.org/10.1037/0278-7393.34.3.514.

SENN, S. (2007). *Statistical Issues in Drug Development*, 2nd ed. Wiley, Chichester, UK.

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 721–743
https://doi.org/10.1214/21-AOAS1508

SCALABLE CHANGE-POINT AND ANOMALY DETECTION IN CROSS-CORRELATED DATA WITH AN APPLICATION TO CONDITION MONITORING

BY MARTIN TVETEN[1,a], IDRIS A. ECKLEY[2,b] AND PAUL FEARNHEAD[2,c]

[1]*Department of Mathematics, University of Oslo,* [a]*tveten@nr.no*
[2]*Mathematics and Statistics, Lancaster University,* [b]*i.eckley@lancaster.ac.uk,* [c]*p.fearnhead@lancaster.ac.uk*

Motivated by a condition monitoring application arising from subsea engineering, we derive a novel, scalable approach to detecting anomalous mean structure in a subset of correlated multivariate time series. Given the need to analyse such series efficiently, we explore a computationally efficient approximation of the maximum likelihood solution to the resulting modelling framework and develop a new dynamic programming algorithm for solving the resulting binary quadratic programme when the precision matrix of the time series at any given time point is banded. Through a comprehensive simulation study we show that the resulting methods perform favorably compared to competing methods, both in the anomaly and change detection settings, even when the sparsity structure of the precision matrix estimate is misspecified. We also demonstrate its ability to correctly detect faulty time periods of a pump within the motivating application.

REFERENCES

BARDWELL, L., FEARNHEAD, P., ECKLEY, I. A., SMITH, S. and SPOTT, M. (2019). Most recent changepoint detection in panel data. *Technometrics* **61** 88–98. MR3933661 https://doi.org/10.1080/00401706.2018.1438926

BHATTACHARJEE, M., BANERJEE, M. and MICHAILIDIS, G. (2019). Change point estimation in panel data with temporal and cross-sectional dependence. Preprint. Available at arXiv:1904.11101.

BLEAKLEY, K. and VERT, J.-P. (2011). The group fused Lasso for multiple change-point detection. Preprint. Available at arXiv:1106.4199.

CHO, H. (2016). Change-point detection in panel data via double CUSUM statistic. *Electron. J. Stat.* **10** 2000–2038. MR3522667 https://doi.org/10.1214/16-EJS1155

CHO, H. and FRYZLEWICZ, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 475–507. MR3310536 https://doi.org/10.1111/rssb.12079

CUTHILL, E. and MCKEE, J. (1969). Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the* 1969 24*th National Conference. ACM '*69 157–172. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/800195.805928

EGUSQUIZA, E., VALERO, C., VALENTIN, D., PRESAS, A. and RODRIGUEZ, C. G. (2015). Condition monitoring of pump-turbines. New challenges. *Measurement* **67** 151–163. https://doi.org/10.1016/j.measurement.2015.01.004

FEARNHEAD, P. and RIGAILL, G. (2019). Changepoint detection in the presence of outliers. *J. Amer. Statist. Assoc.* **114** 169–183. MR3941246 https://doi.org/10.1080/01621459.2017.1385466

FISCH, A. T. M., BARDWELL, L. and ECKLEY, I. A. (2020). Real time anomaly detection and categorisation. Preprint. Available at arXiv:2009.06670.

FISCH, A. T. M., ECKLEY, I. A. and FEARNHEAD, P. (2021a). A linear time method for the detection of point and collective anomalies. *Stat. Anal. Data Min.* To appear. Available at arXiv:1806.01947.

FISCH, A. T. M., ECKLEY, I. A. and FEARNHEAD, P. (2021b). Subset multivariate collective and point anomaly detection. *J. Comput. Graph. Statist.* 1–31. https://doi.org/10.1080/10618600.2021.1987257

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441. https://doi.org/10.1093/biostatistics/kxm045

Key words and phrases. Anomaly, binary quadratic programme, change points, cross-correlation, outliers.

FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42** 2243–2281. MR3269979 https://doi.org/10.1214/14-AOS1245

GAREY, M. R. and JOHNSON, D. S. (1979). *Computers and Intractability*: *A Guide to the Theory of NP-Completeness*. W. H. Freeman and Co., San Francisco, CA. MR0519066

HENRIQUEZ, P., ALONSO, J. B., FERRER, M. A. and TRAVIESO, C. M. (2014). Review of automatic fault diagnosis systems using audio and vibration signals. *IEEE Trans. Syst. Man Cybern. Syst.* **44** 642–652. https://doi.org/10.1109/TSMCC.2013.2257752

HORVÁTH, L. and HUŠKOVÁ, M. (2012). Change-point detection in panel data. *J. Time Series Anal.* **33** 631–648. MR2944843 https://doi.org/10.1111/j.1467-9892.2012.00796.x

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218. https://doi.org/10.1007/BF01908075

JENG, X. J., CAI, T. T. and LI, H. (2013). Simultaneous discovery of rare and common segment variants. *Biometrika* **100** 157–172. MR3034330 https://doi.org/10.1093/biomet/ass059

JIRAK, M. (2015). Uniform change point tests in high dimension. *Ann. Statist.* **43** 2451–2483. MR3405600 https://doi.org/10.1214/15-AOS1347

KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418 https://doi.org/10.1080/01621459.2012.737745

KIRCH, C., MUHSAL, B. and OMBAO, H. (2015). Detection of changes in multivariate time series with application to EEG data. *J. Amer. Statist. Assoc.* **110** 1197–1216. MR3420695 https://doi.org/10.1080/01621459.2014.957545

KLANDERMAN, M. C., NEWHART, K. B., CATH, T. Y. and HERING, A. S. (2020). Fault isolation for a complex decentralized waste water treatment facility. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 931–951. MR4133153 https://doi.org/10.1111/rssc.12429

KOVÁCS, S., LI, H., BÜHLMANN, P. and MUNK, A. (2020). Seeded binary segmentation: A general methodology for fast and optimal change point detection. Preprint. Available at arXiv:2002.06633.

LEWIS, J. G. (1982). Algorithm 582: The Gibbs–Poole–Stockmeyer and Gibbs–King algorithms for reordering sparse matrices. *ACM Trans. Math. Software* **8** 190–194. https://doi.org/10.1145/355993.355999

LI, J., XU, M., ZHONG, P.-S. and LI, L. (2019). Change point detection in the mean of high-dimensional time series data under dependence. Preprint. Available at arXiv:1903.07006.

LIU, H., GAO, C. and SAMWORTH, R. J. (2021). Minimax rates in sparse, high-dimensional change point detection. *Ann. Statist.* **49** 1081–1112. MR4255120 https://doi.org/10.1214/20-aos1994

ÖLLERER, V. and CROUX, C. (2015). Robust high-dimensional precision matrix estimation. In *Modern Nonparametric*, *Robust and Multivariate Methods* (K. Nordhausen and S. Taskinen, eds.) 325–350. Springer, Cham. MR3444335

SAFIKHANI, A. and SHOJAIE, A. (2020). Joint structural break detection and parameter estimation in high-dimensional nonstationary VAR models. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2020.1770097

SUSTIK, M. A. and CALDERHEAD, B. (2012). GLASSOFAST: An efficient GLASSO implementation. UTCS Technical Report **TR-12-29**.

TCHAKOUA, P., WAMKEUE, R., OUHROUCHE, M., SLAOUI-HASNAOUI, F., TAMEGHE, T. A. and EKEMB, G. (2014). Wind turbine condition monitoring: State-of-the-art review, new trends, and future challenges. *Energies* **7** 2595–2630. https://doi.org/10.3390/en7042595

TVETEN, M., ECKLEY, I. A. and FEARNHEAD, P. (2022). Supplement to "Scalable change-point and anomaly detection in cross-correlated data with an application to condition monitoring." https://doi.org/10.1214/21-AOAS1508SUPPA, https://doi.org/10.1214/21-AOAS1508SUPPB

VER HOEF, J. M., HANKS, E. M. and HOOTEN, M. B. (2018). On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models. *Spat. Stat.* **25** 68–85. MR3809256 https://doi.org/10.1016/j.spasta.2018.04.006

WANG, T. and SAMWORTH, R. J. (2018). High dimensional change point estimation via sparse projection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 57–83. MR3744712 https://doi.org/10.1111/rssb.12243

WANG, D., YU, Y., RINALDO, A. and WILLETT, R. (2020). Localizing changes in high-dimensional vector autoregressive processes. Preprint. Available at arXiv:1909.06359.

WESTERLUND, J. (2019). Common breaks in means for cross-correlated fixed-T panel data. *J. Time Series Anal.* **40** 248–255. MR3915529 https://doi.org/10.1111/jtsa.12407

XIE, Y. and SIEGMUND, D. (2013). Sequential multi-sensor change-point detection. *Ann. Statist.* **41** 670–692. MR3099117 https://doi.org/10.1214/13-AOS1094

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 744–764
https://doi.org/10.1214/21-AOAS1512

ADAPTIVE DESIGN FOR GAUSSIAN PROCESS REGRESSION UNDER CENSORING

BY JIALEI CHEN[1,a], SIMON MAK[2,d], V. ROSHAN JOSEPH[1,b] AND CHUCK ZHANG[1,c]

[1]*H. Milton Stewart School of Industrial & Systems Engineering Georgia Institute of Technology,* [a]*jialei.chen@gatech.edu,* [b]*roshan@gatech.edu,* [c]*chuck.zhang@gatech.edu*

[2]*Department of Statistical Science, Duke University,* [d]*sm769@duke.edu*

A key objective in engineering problems is to predict an unknown experimental surface over an input domain. In complex physical experiments this may be hampered by response censoring which results in a significant loss of information. For such problems, experimental design is paramount for maximizing predictive power using a small number of expensive experimental runs. To tackle this, we propose a novel adaptive design method, called the integrated *censored* mean-squared error (ICMSE) method. The ICMSE method first estimates the posterior probability of a new observation being censored, then adaptively chooses design points that minimize predictive uncertainty under censoring. Adopting a Gaussian process regression model with product correlation function, the proposed ICMSE criterion is easy to evaluate which allows for efficient design optimization. We demonstrate the effectiveness of the ICMSE design in two real-world applications on surgical planning and wafer manufacturing.

REFERENCES

ANKENMAN, B., NELSON, B. L. and STAUM, J. (2010). Stochastic kriging for simulation metamodeling. *Oper. Res.* **58** 371–382. MR2674803 https://doi.org/10.1287/opre.1090.0754

BECT, J., BACHOC, F. and GINSBOURGER, D. (2019). A supermartingale approach to Gaussian process based sequential design of experiments. *Bernoulli* **25** 2883–2919. MR4003568 https://doi.org/10.3150/18-BEJ1074

BINOIS, M., HUANG, J., GRAMACY, R. B. and LUDKOVSKI, M. (2019). Replication or exploration? Sequential design for stochastic simulation experiments. *Technometrics* **61** 7–23. MR3933655 https://doi.org/10.1080/00401706.2018.1469433

BORTH, D. M. (1996). Optimal experimental designs for (possibly) censored data. *Chemom. Intell. Lab. Syst.* **32** 25–35.

BROOKS, R. J. (1982). On the loss of information through censoring. *Biometrika* **69** 137–144. MR0655678 https://doi.org/10.1093/biomet/69.1.137

CAO, F., BA, S., BRENNEMAN, W. A. and JOSEPH, V. R. (2018). Model calibration with censored data. *Technometrics* **60** 255–262. MR3804253 https://doi.org/10.1080/00401706.2017.1345704

CHEN, R.-B., WANG, W. and WU, C. F. J. (2017). Sequential designs based on Bayesian uncertainty quantification in sparse representation surrogate modeling. *Technometrics* **59** 139–152. MR3635038 https://doi.org/10.1080/00401706.2016.1172027

CHEN, J., WANG, K., ZHANG, C. and WANG, B. (2018a). An efficient statistical approach to design 3D-printed metamaterials for mimicking mechanical properties of soft biological tissues. *Addit. Manuf.* **24** 341–352.

CHEN, J., XIE, Y., WANG, K., WANG, Z. H., LAHOTI, G., ZHANG, C., VANNAN, M. A., WANG, B. and QIAN, Z. (2018b). Generative invertible networks (GIN): Pathophysiology-interpretable feature mapping and virtual patient generation. arXiv preprint. Available at arXiv:1808.04495.

CHEN, J., MAK, S., JOSEPH, V. R. and ZHANG, C. (2021). Function-on-function kriging, with applications to 3D printing of aortic tissues. *Technometrics* **63** 384–395.

CHEN, J., MAK, S., JOSEPH, V. R and ZHANG, C. (2022). Supplement to "Adaptive design for Gaussian process regression under censoring." https://doi.org/10.1214/21-AOAS1512SUPP

THERMO ELECTRIC COMPANY (2010). Wafer sensors. Available at http://www.te-direct.com/products/silicon-wafers/.

Key words and phrases. Adaptive sampling, censored experiments, experimental design, kriging, multifidelity modeling.

DA VEIGA, S. and MARREL, A. (2012). Gaussian process modeling with inequality constraints. *Ann. Fac. Sci. Toulouse Math.* (6) **21** 529–555. MR3076411 https://doi.org/10.5802/afst.1344

DICKINSON, E. J., EKSTRÖM, H. and FONTES, E. (2014). COMSOL multiphysics®: Finite element software for electrochemical analysis. A mini-review. *Electrochem. Commun.* **40** 71–74.

DING, L., MAK, S. and WU, C. F. J. (2019). Bdrygp: A new Gaussian process model for incorporating boundary information. arXiv preprint. Available at arXiv:1908.08868.

FETEIRA, A. (2009). Negative temperature coefficient resistance (NTCR) ceramic thermistors: An industrial perspective. *J. Amer. Ceram. Soc.* **92** 967–983.

GIBSON, I., ROSEN, D. W. and STUCKER, B. (2014). *Additive Manufacturing Technologies* **17**. Springer, Berlin.

GINSBOURGER, D., LE RICHE, R. and CARRARO, L. (2010). Kriging is well-suited to parallelize optimization. In *Computational Intelligence in Expensive Optimization Problems* 131–162. Springer, Berlin.

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

GOODSON, K., FLIK, M., SU, L. and ANTONIADIS, D. A. (1993). Annealing-temperature dependence of the thermal conductivity of lpcvd silicon-dioxide layers. *IEEE Electron Device Lett.* **14** 490–492.

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

GROOT, P., LUCAS, P., CANO, A., GÓMEZ-OLMEDO, M. and NIELSEN, T. (2012). Gaussian process regression with censored data using expectation propagation. In *PGM* 2012: *Proceedings of the Sixth European Workshop on Probabilistic Graphical Models*, *PGM'*12 (A. Cano, M. Gómez-Olmedo and T. D. Nielsen, eds.) 115–122. DECSAI, Granada.

HENKENJOHANN, N., GÖBEL, R., KLEINER, M. and KUNERT, J. (2005). An adaptive sequential procedure for efficient optimization of the sheet metal spinning process. *Qual. Reliab. Eng. Int.* **21** 439–455.

JIN, R., CHANG, C.-J. and SHI, J. (2012). Sequential measurement strategy for wafer geometric profile estimation. *IIE Trans.* **44** 1–12.

JOHNSON, M. E., MOORE, L. M. and YLVISAKER, D. (1990). Minimax and maximin distance designs. *J. Statist. Plann. Inference* **26** 131–148. MR1079258 https://doi.org/10.1016/0378-3758(90)90122-B

JOSEPH, V. R. (2016). Rejoinder. *Qual. Eng.* **28** 42–44.

JOSEPH, V. R., GUL, E. and BA, S. (2015). Maximum projection designs for computer experiments. *Biometrika* **102** 371–380. MR3371010 https://doi.org/10.1093/biomet/asv002

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

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

LAM, C. Q. (2008). *Sequential Adaptive Designs in Computer Experiments for Response Surface Model Fit*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—The Ohio State University. MR2711949

LIAO, K., SCHULTESIZ, C. R., HUNSTON, D. L. and BRINSON, L. C. (1998). Long-term durability of fiber-reinforced polymer-matrix composite materials for infrastructure applications: A review. *J. Adv. Mater.* **30** 3–40.

LOEPPKY, J. L., SACKS, J. and WELCH, W. J. (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics* **51** 366–376. MR2756473 https://doi.org/10.1198/TECH.2009.08040

LÓPEZ-LOPERA, A. F., BACHOC, F., DURRANDE, N. and ROUSTANT, O. (2018). Finite-dimensional Gaussian approximation with linear inequality constraints. *SIAM/ASA J. Uncertain. Quantificat.* **6** 1224–1255. MR3857898 https://doi.org/10.1137/17M1153157

MAK, S., SUNG, C.-L., WANG, X., YEH, S.-T., CHANG, Y.-H., JOSEPH, V. R., YANG, V. and WU, C. F. J. (2018). An efficient surrogate model for emulation and physics extraction of large eddy simulations. *J. Amer. Statist. Assoc.* **113** 1443–1456. MR3902221 https://doi.org/10.1080/01621459.2017.1409123

MATHÉRON, G. (1963). Principles of geostatistics. *Econ. Geol.* **58** 1246–1266.

MONROE, E. M. and PAN, R. (2008). Experimental design considerations for accelerated life tests with nonlinear constraints and censoring. *J. Qual. Technol.* **40** 355–367.

MORRIS, M. D. and MITCHELL, T. J. (1995). Exploratory designs for computational experiments. *J. Statist. Plann. Inference* **43** 381–402.

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

QIAN, Z., WANG, K., LIU, S., ZHOU, X., RAJAGOPAL, V., MEDURI, C., KAUTEN, J. R., CHANG, Y.-H., WU, C. et al. (2017). Quantitative prediction of paravalvular leak in transcatheter aortic valve replacement based on tissue-mimicking 3D printing. *JACC Cardiovasc. Imaging* **10** 719–731.

QUIRK, M. and SERDA, J. (2001). *Semiconductor Manufacturing Technology* **1**. Prentice Hall, Upper Saddle River, NJ.

RENGIER, F., MEHNDIRATTA, A., VON TENGG-KOBLIGK, H., ZECHMANN, C. M., UNTERHINNINGHOFEN, R., KAUCZOR, H.-U. and GIESEL, F. L. (2010). 3D printing based on imaging data: Review of medical applications. *Internat. J. Comput. Assisted Radiol. Surg.* **5** 335–341.

SACKS, J., SCHILLER, S. B. and WELCH, W. J. (1989). Designs for computer experiments. *Technometrics* **31** 41–47. MR0997669 https://doi.org/10.2307/1270363

SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. MR1041765

SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2018). *The Design and Analysis of Computer Experiments*. *Springer Series in Statistics*. Springer, New York. MR3887662

SHEWRY, M. C. and WYNN, H. P. (1987). Maximum entropy sampling. *J. Appl. Stat.* **14** 165–170.

SICARD, D., HAAK, A. J., CHOI, K. M., CRAIG, A. R., FREDENBURGH, L. E. and TSCHUMPERLIN, D. J. (2018). Aging and anatomical variations in lung tissue stiffness. *Am. J. Physiol., Lung Cell. Mol. Physiol.* **314** L946–L955. https://doi.org/10.1152/ajplung.00415.2017

SINGH, R., FAKHRUDDIN, M. and POOLE, K. (2000). Rapid photothermal processing as a semiconductor manufacturing technology for the 21st century. *Appl. Surf. Sci.* **168** 198–203.

SOBOL', I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *Ž. Vyčisl. Mat. Mat. Fiz.* **7** 784–802.

VAN GURP, M. and PALMEN, J. (1998). Time-temperature superposition for polymeric blends. *Rheol. Bull.* **67** 5–8.

WANG, K., WU, C., QIAN, Z., ZHANG, C., WANG, B. and VANNAN, M. A. (2016). Dual-material 3D printed metamaterials with tunable mechanical properties for patient-specific tissue-mimicking phantoms. *Addit. Manuf.* **12** 31–37.

WILHELM, S. and MANJUNATH, B. G. (2010). tmvtnorm: A package for the truncated multivariate normal distribution. *R J.* **2** 25–29.

XIONG, S., QIAN, P. Z. G. and WU, C. F. J. (2013). Sequential design and analysis of high-accuracy and low-accuracy computer codes. *Technometrics* **55** 37–46. MR3038483 https://doi.org/10.1080/00401706.2012.723572

YPMA, J., BORCHERS, H. and EDDELBUETTEL, D. (2014). nloptr: R interface to nlopt. *R J.* **1**.

ZIENKIEWICZ, O. C., TAYLOR, R. L., ZIENKIEWICZ, O. C. and TAYLOR, R. L. (1977). *The Finite Element Method* **36**. McGraw-Hill, London.

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 765–790
https://doi.org/10.1214/21-AOAS1515

COMPOSITE MIXTURE OF LOG-LINEAR MODELS WITH APPLICATION TO PSYCHIATRIC STUDIES

BY EMANUELE ALIVERTI[1,a] AND DAVID B. DUNSON[2,b]

[1]*Department of Economics, University Ca' Foscari Venezia,* [a]*emanuele.aliverti@unive.it*
[2]*Department of Statistical Sciences, Duke University,* [b]*dunson@duke.edu*

Psychiatric studies of suicide provide fundamental insights on the evolution of severe psychopathologies, and contribute to the development of early treatment interventions. Our focus is on modelling different traits of psychosis and their interconnections, focusing on a case study on suicide attempt survivors. Such aspects are recorded via multivariate categorical data, involving a large numbers of items for multiple subjects. Current methods for multivariate categorical data—such as penalized log-linear models and latent structure analysis—are either limited to low-dimensional settings or include parameters with difficult interpretation. Motivated by this application, this article proposes a new class of approaches, which we refer to as Mixture of Log Linear models (MILLS). Combining latent class analysis and log-linear models, MILLS defines a novel Bayesian approach to model complex multivariate categorical data with flexibility and interpretability, providing interesting insights on the relationship between psychotic diseases and psychological aspects in suicide attempt survivors.

REFERENCES

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

AIROLDI, E. M., BLEI, D., EROSHEVA, E. A. and FIENBERG, S. E. (2014). *Handbook of Mixed Membership Models and Their Applications*. CRC press.

ANDERSEN, E. B. (1982). Latent structure analysis: A survey. *Scand. J. Stat.* **9** 1–12. MR0651855

BERGSMA, W. P. and RUDAS, T. (2002). Marginal models for categorical data. *Ann. Statist.* **30** 140–159. MR1892659 https://doi.org/10.1214/aos/1015362188

BHATTACHARYA, A. and DUNSON, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *J. Amer. Statist. Assoc.* **107** 362–377. MR2949366 https://doi.org/10.1080/01621459.2011.646934

BRADLEY, J. R., HOLAN, S. H. and WIKLE, C. K. (2020). Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family. *J. Amer. Statist. Assoc.* **115** 2037–2052. MR4189775 https://doi.org/10.1080/01621459.2019.1677471

BRITTON, P. C., BOHNERT, A. S., WINES JR., J. D. and CONNER, K. R. (2012). A procedure that differentiates unintentional from intentional overdose in opioid abusers. *Addictive Behaviors* **37** 127–130.

CHANG, I. H. and MUKERJEE, R. (2006). Probability matching property of adjusted likelihoods. *Statist. Probab. Lett.* **76** 838–842. MR2266098 https://doi.org/10.1016/j.spl.2005.10.015

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

CUSI, A. M., MACQUEEN, G. M., SPRENG, R. N. and MCKINNON, M. C. (2011). Altered empathic responding in major depressive disorder: Relation to symptom severity, illness burden, and psychosocial outcome. *Psychiatry Res.* **188** 231–236.

DAVIS, M. H. (1980). A muntidimensional approach to individual differences in empathy. *JSAS Catalogue of Selected Documents in Psychology* **10**.

DEROGATIS, L. R., LIPMAN, R. S. and COVI, L. (1973). SCL-90: An outpatient psychiatric rating scale–preliminary report. *Psychopharmacol Bull.* **9** 13–28.

DE BEURS, D., FRIED, E. I., WETHERALL, K., CLEARE, S., O'CONNOR, D. B., FERGUSON, E., O'CARROLL, R. E. and O'CONNOR, R. C. (2019). Exploring the psychology of suicidal ideation: A theory driven network analysis. *Behaviour Research and Therapy* **120** 103419.

Key words and phrases. Psychiatric profiles, Bayesian modelling, categorical data, contingency table, log-linear models, mixture model.

De Leo, D., Burgis, S., Bertolote, J. M., Kerkhof, A. and Bille-Brahe, U. (2004). Definitions of suicidal behaviour. *Suicidal Behaviour*: *Theories and Research Findings* 17–39.

Dobra, A. and Massam, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. *Stat. Methodol.* **7** 240–253. MR2643600 https://doi.org/10.1016/j.stamet.2009.04.002

Dobra, A. and Mohammadi, R. (2018). Loglinear model selection and human mobility. *Ann. Appl. Stat.* **12** 815–845. MR3834287 https://doi.org/10.1214/18-AOAS1164

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

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80** 3–26. MR1225211 https://doi.org/10.1093/biomet/80.1.3

Erosheva, E. A. (2005). Comparing latent structures of the grade of membership, Rasch, and latent class models. *Psychometrika* **70** 619–628. MR2272507 https://doi.org/10.1007/s11336-001-0899-y

Ferrari, D., Qian, G. and Hunter, T. (2016). Parsimonious and efficient likelihood composition by Gibbs sampling. *J. Comput. Graph. Statist.* **25** 935–953. MR3533646 https://doi.org/10.1080/10618600.2015.1058799

Fienberg, S. E. and Rinaldo, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *J. Statist. Plann. Inference* **137** 3430–3445. MR2363267 https://doi.org/10.1016/j.jspi.2007.03.022

Fontenelle, L. F., Soares, I. D., Miele, F., Borges, M. C., Prazeres, A. M., Rangé, B. P. and Moll, J. (2009). Empathy and symptoms dimensions of patients with obsessive–compulsive disorder. *J. Psychiatr. Res.* **43** 455–463.

Fraser, D. A. S. and Reid, N. (2020). Combining likelihood and significance functions. *Statist. Sinica* **30** 1–15. MR4285482 https://doi.org/10.5705/ss.202016.0508

Frühwirth-Schnatter, S., Celeux, G. and Robert, C. P., eds. (2019). *Handbook of Mixture Analysis*. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR3889980

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

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. Ser. B* **54** 657–699. MR1185217

Gilet, A.-L., Mella, N., Studer, J., Grühn, D. and Labouvie-Vief, G. (2013). Assessing dispositional empathy in adults: A French validation of the Interpersonal Reactivity Index (IRI). *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement* **45** 42.

Goodfellow, B., Kõlves, K. and de Leo, D. (2019). Contemporary definitions of suicidal behavior: A systematic literature review. *Suicide Life Threat. Behav/* **49** 488–504. https://doi.org/10.1111/sltb.12457

Greco, L., Racugno, W. and Ventura, L. (2008). Robust likelihood functions in Bayesian inference. *J. Statist. Plann. Inference* **138** 1258–1270. MR2388009 https://doi.org/10.1016/j.jspi.2007.05.001

Guttman, H. and Laporte, L. (2002). Alexithymia, empathy, and psychological symptoms in a family context. *Comprehensive Psychiatry* **43** 448–455.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity*: *The Lasso and Generalizations*. *Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. MR3616141

Hawton, K. and Fagg, J. (1988). Suicide, and other causes of death, following attempted suicide. *Br. J. Psychiatry* **152** 359–366.

Huang, Z. and Ferrari, D. (2017). Fast construction of efficient composite likelihood equations. Preprint. Available at arXiv:1709.03234.

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

Johndrow, J. and Bhattacharya, A. (2018). Optimal Gaussian approximations to the posterior for log-linear models with Diaconis–Ylvisaker priors. *Bayesian Anal.* **13** 201–223. MR3737949 https://doi.org/10.1214/16-BA1046

Johndrow, J. E., Bhattacharya, A. and Dunson, D. B. (2017). Tensor decompositions and sparse log-linear models. *Ann. Statist.* **45** 1–38. MR3611485 https://doi.org/10.1214/15-AOS1414

Kelleher, I., Harley, M., Murtagh, A. and Cannon, M. (2011). Are screening instruments valid for psychotic-like experiences? A validation study of screening questions for psychotic-like experiences using in-depth clinical interview. *Schizophr. Bull.* **37** 362–369.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056 https://doi.org/10.1137/07070111X

Ladisich, W. and Feil, W. (1988). Empathy in psychiatric patients. *British Journal of Medical Psychology* **61** 155–162.

LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. The Clarendon Press, New York. MR1419991

LAWLEY, D. N. (1943). On problems connected with item selection and test construction. *Proc. Roy. Soc. Edinburgh Sect. A* **61** 273–287. MR0007966

LAZAR, N. A. (2003). Bayesian empirical likelihood. *Biometrika* **90** 319–326. MR1986649 https://doi.org/10.1093/biomet/90.2.319

LAZARSFELD, P. F. (1950). The logical and mathematical foundation of latent structure analysis. *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction* 362–412.

LEGRAMANTI, S., DURANTE, D. and DUNSON, D. B. (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika* **107** 745–752. MR4138988 https://doi.org/10.1093/biomet/asaa008

LETAC, G. and MASSAM, H. (2012). Bayes factors and the geometry of discrete hierarchical loglinear models. *Ann. Statist.* **40** 861–890. MR2985936 https://doi.org/10.1214/12-AOS974

LINDSAY, B. G., YI, G. Y. and SUN, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statist. Sinica* **21** 71–105. MR2796854

LUPPARELLI, M., MARCHETTI, G. M. and BERGSMA, W. P. (2009). Parameterizations and fitting of bi-directed graph models to categorical data. *Scand. J. Stat.* **36** 559–576. MR2549710 https://doi.org/10.1111/j.1467-9469.2008.00638.x

MANRIQUE-VALLIER, D. (2014). Longitudinal mixed membership trajectory models for disability survey data. *Ann. Appl. Stat.* **8** 2268–2291. MR3292497 https://doi.org/10.1214/14-AOAS769

MARDIA, K. V., KENT, J. T., HUGHES, G. and TAYLOR, C. C. (2009). Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika* **96** 975–982. MR2767282 https://doi.org/10.1093/biomet/asp056

MASSAM, H., LIU, J. and DOBRA, A. (2009). A conjugate prior for discrete hierarchical log-linear models. *Ann. Statist.* **37** 3431–3467. MR2549565 https://doi.org/10.1214/08-AOS669

MASSAM, H. and WANG, N. (2018). Local conditional and marginal approach to parameter estimation in discrete graphical models. *J. Multivariate Anal.* **164** 1–21. MR3738130 https://doi.org/10.1016/j.jmva.2017.10.003

MCCORMICK, L. M., BRUMM, M. C., BEADLE, J. N., PARADISO, S., YAMADA, T. and ANDREASEN, N. (2012). Mirror neuron function, psychosis, and empathy in schizophrenia. *Psychiatry Research. Neuroimaging* **201** 233–239.

MCHUGH, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika* **21** 331–347. MR0082427 https://doi.org/10.1007/BF02296300

MENG, Z., WEI, D., WIESEL, A. and HERO III, A. (2013). Distributed learning of Gaussian graphical models via marginal likelihoods. In *Artificial Intelligence and Statistics* 39–47.

MILLER, J. W. (2019). Asymptotic normality, concentration, and coverage of generalized posteriors. Preprint. Available at arXiv:1907.09611.

NARDI, Y. and RINALDO, A. (2012). The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli* **18** 945–974. MR2948908 https://doi.org/10.3150/11-BEJ364

NOCK, M. K., BORGES, G., BROMET, E. J., ALONSO, J., ANGERMEYER, M., BEAUTRAIS, A., BRUFFAERTS, R., CHIU, W. T., DE GIROLAMO, G. et al. (2008). Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *Br. J. Psychiatry* **192** 98–105.

PACE, L., SALVAN, A. and SARTORI, N. (2019). Efficient composite likelihood for a scalar parameter of interest. *Stat.* **8** e222. MR3938285 https://doi.org/10.1002/sta4.222

PAULI, F., RACUGNO, W. and VENTURA, L. (2011). Bayesian composite marginal likelihoods. *Statist. Sinica* **21** 149–164. MR2796857

PERRONE-MCGOVERN, K. M., OLIVEIRA-SILVA, P., SIMON-DACK, S., LEFDAHL-DAVIS, E., ADAMS, D., MCCONNELL, J., HOWELL, D., HESS, R., DAVIS, A. et al. (2014). Effects of empathy and conflict resolution strategies on psychophysiological arousal and satisfaction in romantic relationships. *Applied Psychophysiology and Biofeedback* **39** 19–25.

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

PRINZ, U., NUTZINGER, D. O., SCHULZ, H., PETERMANN, F., BRAUKHAUS, C. and ANDREAS, S. (2013). Comparative psychometric analyses of the SCL-90-R and its short versions in patients with affective disorders. *BMC Psychiatry* **13** 104. https://doi.org/10.1186/1471-244X-13-104

PRUNAS, A., SARNO, I., PRETI, E., MADEDDU, F. and PERUGINI, M. (2012). Psychometric properties of the Italian version of the scl-90-r: A study on a large community sample. *European Psychiatry* **27** 591–597.

RAFTERY, A. E. (1985). A model for high-order Markov chains. *J. Roy. Statist. Soc. Ser. B* **47** 528–539. MR0844484

RAFTERY, A. E., MADIGAN, D. and VOLINSKY, C. T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance. In *Bayesian Statistics*, 5 (*Alicante*, 1994). *Oxford Sci. Publ.* 323–349. Oxford Univ. Press, New York. MR1425413

RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343 https://doi.org/10.1214/09-AOS691

RIBATET, M., COOLEY, D. and DAVISON, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statist. Sinica* **22** 813–845. MR2954363

ROUSSEAU, J. and MENGERSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 689–710. MR2867454 https://doi.org/10.1111/j.1467-9868.2011.00781.x

ROVERATO, A., LUPPARELLI, M. and LA ROCCA, L. (2013). Log-mean linear models for binary data. *Biometrika* **100** 485–494. MR3068448 https://doi.org/10.1093/biomet/ass080

RUSSO, M., DURANTE, D. and SCARPA, B. (2018). Bayesian inference on group differences in multivariate categorical data. *Comput. Statist. Data Anal.* **126** 136–149. MR3808395 https://doi.org/10.1016/j.csda.2018.04.010

SCHREITER, S., PIJNENBORG, G. and AAN HET ROT, M. (2013). Empathy in adults with clinical or subclinical depressive symptoms. *J. Affective Disorders* **150** 1–16.

SCOCCO, P. and DE LEO, D. (2002). One-year prevalence of death thoughts, suicide ideation and behaviours in an elderly population. *International Journal of Geriatric Psychiatry* **17** 842–846.

SCOCCO, P., ALIVERTI, E., TOFFOL, E., ANDRETTA, G. and CAPIZZI, G. (2020). Empathy profiles differ by gender in people who have and have not attempted suicide. *Journal of Affective Disorders Reports* **2** 100024.

SNIJDERS, T. A. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *J. Soc. Struct.* **3** 1–40.

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

WANG, W., ZHOU, Y., WANG, J., XU, H., WEI, S., WANG, D., WANG, L. and ZHANG, X. (2020). Prevalence, clinical correlates of suicide attempt and its relationship with empathy in patients with schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 109863.

ZHANG, K., SZANTO, K., CLARK, L. and DOMBROVSKI, A. Y. (2019). Behavioral empathy failures and suicidal behavior. *Behav. Res. Ther.* **120** 103329. https://doi.org/10.1016/j.brat.2018.10.019

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 791–815
https://doi.org/10.1214/21-AOAS1519

INHOMOGENEOUS SPATIO-TEMPORAL POINT PROCESSES ON LINEAR NETWORKS FOR VISITORS' STOPS DATA

BY NICOLETTA D'ANGELO[1,a], GIADA ADELFIO[1,b], ANTONINO ABBRUZZO[1,c] AND JORGE MATEU[2,d]

[1]*Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo,* [a]*nicoletta.dangelo@unipa.it,* [b]*giada.adelfio@unipa.it,* [c]*antonino.abbruzzo@unipa.it*

[2]*Department of Mathematics, University Jaume I,* [d]*mateu@uji.es*

We analyse the spatio-temporal distribution of visitors' stops by touristic attractions in Palermo (Italy), using theory of stochastic point processes living on linear networks. We first propose an inhomogeneous Poisson point process model with a separable parametric spatio-temporal first-order intensity. We account for the spatial interaction among points on the given network, fitting a Gibbs point process model with mixed effects for the purely spatial component. This allows us to study first-order and second-order properties of the point pattern, accounting both for the spatio-temporal clustering and interaction and for the spatio-temporal scale at which they operate. Due to the strong degree of clustering in the data, we then formulate a more complex model, fitting a spatio-temporal log-Gaussian Cox process to the point process on the linear network, addressing the problem of the choice of the most appropriate distance metric.

REFERENCES

ABBRUZZO, A., FERRANTE, M. and DE CANTIS, S. (2021). A pre-processing and network analysis of GPS tracking data. *Spatial Econ. Anal.* **16** 217–240.

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

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

BADDELEY, A., BÁRÁNY, I. and SCHNEIDER, R. (2006). *Stochastic Geometry*: *Lectures Given at the CIME Summer School Held in Martina Franca*, *Italy*, *September* 13–18, 2004. Springer.

BADDELEY, A. J., MØLLER, J. and WAAGEPETERSEN, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Stat. Neerl.* **54** 329–350. MR1804002 https://doi.org/10.1111/1467-9574.00144

BADDELEY, A., CHANG, Y.-M., SONG, Y. and TURNER, R. (2012). Nonparametric estimation of the dependence of a spatial point process on spatial covariates. *Stat. Interface* **5** 221–236. MR2928072 https://doi.org/10.4310/SII.2012.v5.n2.a7

BADDELEY, A., NAIR, G., RAKSHIT, S. and MCSWIGGAN, G. (2017). "Stationary" point processes are uncommon on linear networks. *Stat* **6** 68–78. MR3613182 https://doi.org/10.1002/sta4.135

BADDELEY, A., NAIR, G., RAKSHIT, S., MCSWIGGAN, G. and DAVIES, T. M. (2021). Analysing point patterns on networks—a review. *Spat. Stat.* **42** Paper No. 100435, 35. MR4233256 https://doi.org/10.1016/j.spasta.2020.100435

BRIX, A. and DIGGLE, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 823–841. MR1872069 https://doi.org/10.1111/1467-9868.00315

BUTZ, W. P. and TORREY, B. B. (2006). Some frontiers in social science. *Science* **312** 1898–1900.

CRONIE, O., MORADI, M. and MATEU, J. (2020). Inhomogeneous higher-order summary statistics for point processes on linear networks. *Stat. Comput.* **30** 1221–1239. MR4137248 https://doi.org/10.1007/s11222-020-09942-w

Key words and phrases. Gibbs point processes, global positioning system, intensity estimation, linear networks, log-Gaussian Cox processes, spatio-temporal point processes.

DAVIES, T. M. and HAZELTON, M. L. (2013). Assessing minimum contrast parameter estimation for spatial and spatiotemporal log-Gaussian Cox processes. *Stat. Neerl.* **67** 355–389. MR3127239 https://doi.org/10.1111/stan.12011

DE CANTIS, S., FERRANTE, M., KAHANI, A. and SHOVAL, N. (2016). Cruise passengers' behavior at the destination: Investigation using GPS technology. *Tour. Manag.* **52** 133–150.

DIGGLE, P. J. (2014). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, 3rd ed. *Monographs on Statistics and Applied Probability* **128**. CRC Press, Boca Raton, FL. MR3113855

ELGETHUN, K., FENSKE, R. A., YOST, M. G. and PALCISKO, G. J. (2003). Time-location analysis for exposure assessment studies of children using a novel global positioning system instrument. *Environ. Health Perspect.* **111** 115–122.

ESTER, M., KRIEGEL, H.-P., SANDER, J., XU, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* **96** 226–231.

GABRIEL, E. and DIGGLE, P. J. (2009). Second-order analysis of inhomogeneous spatio-temporal point process data. *Stat. Neerl.* **63** 43–51. MR2656916 https://doi.org/10.1111/j.1467-9574.2008.00407.x

GABRIEL, E., DIGGLE, P. J., ROWLINGSON, B. and RODRIGUEZ-CORTES, F. J. (2021). stpp: Space-time point pattern simulation, visualisation and analysis. R package version 2.0-5.

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

ILLIAN, J. B. and HENDRICHSEN, D. K. (2010). Gibbs point process models with mixed effects. *Environmetrics* **21** 341–353. MR2842247 https://doi.org/10.1002/env.1008

ILLIAN, J., PENTTINEN, A., STOYAN, H. and STOYAN, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns. Statistics in Practice*. Wiley, Chichester. MR2384630

KALLENBERG, O. (1984). An informal guide to the theory of conditioning in point processes. *Int. Stat. Rev.* **52** 151–164. MR0967208 https://doi.org/10.2307/1403098

LI, Z. and WOOD, S. N. (2020). Faster model matrix crossproducts for large generalized linear models with discretized covariates. *Stat. Comput.* **30** 19–25. MR4057468 https://doi.org/10.1007/s11222-019-09864-2

MATEU, J., MORADI, M. and CRONIE, O. (2020a). Spatio-temporal point patterns on linear networks: Pseudo-separable intensity estimation. *Spat. Stat.* **37** 100400, 11. MR4109593 https://doi.org/10.1016/j.spasta.2019.100400

MORADI, M., CRONIE, O. and MATEU, J. (2020b). stlnpp: Spatio-temporal analysis of point patterns on linear networks. R package version 0.3-7.

MCSWIGGAN, G., BADDELEY, A. and NAIR, G. (2017). Kernel density estimation on a linear network. *Scand. J. Stat.* **44** 324–345. MR3658517 https://doi.org/10.1111/sjos.12255

MØLLER, J., SYVERSVEEN, A. R. and WAAGEPETERSEN, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Stat.* **25** 451–482. MR1650019 https://doi.org/10.1111/1467-9469.00115

MORADI, M. M. and MATEU, J. (2020). First- and second-order characteristics of spatio-temporal point processes on linear networks. *J. Comput. Graph. Statist.* **29** 432–443. MR4153171 https://doi.org/10.1080/10618600.2019.1694524

MORADI, M. M., CRONIE, O., RUBAK, E., LACHIEZE-REY, R., MATEU, J. and BADDELEY, A. (2019). Resample-smoothing of Voronoi intensity estimators. *Stat. Comput.* **29** 995–1010. MR3994614 https://doi.org/10.1007/s11222-018-09850-0

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

PADOAN, S. A. and BEVILACQUA, M. (2015). Analysis of random fields using CompRandFld.

R CORE TEAM (2021). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAKSHIT, S., BADDELEY, A. and NAIR, G. (2019). Efficient code for second order analysis of events on a linear network. *J. Stat. Softw.* **90** 1–37.

RAKSHIT, S., NAIR, G. and BADDELEY, A. (2017). Second-order analysis of point patterns on a network using any distance metric. *Spat. Stat.* **22** 129–154. MR3726179 https://doi.org/10.1016/j.spasta.2017.10.002

SHOVAL, N., KAHANI, A., DE CANTIS, S. and FERRANTE, M. (2020). Impact of incentives on tourist activity in space-time. *Ann. Tour. Res.* **80** 102846.

SIINO, M., ADELFIO, G. and MATEU, J. (2018). Joint second-order parameter estimation for spatio-temporal log-Gaussian Cox processes. *Stoch. Environ. Res. Risk Assess.* **32** 3525–3539.

SILVERMAN, B. W. (2018). *Density Estimation for Statistics and Data Analysis*. Routledge. MR0848134

TAMAYO-URIA, I., MATEU, J. and DIGGLE, P. J. (2014). Modelling of the spatio-temporal distribution of rat sightings in an urban environment. *Spat. Stat.* **9** 192–206. MR3326839 https://doi.org/10.1016/j.spasta.2014.03.005

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

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

WOOD, S. N., GOUDE, Y. and SHAW, S. (2015). Generalized additive models for large data sets. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 139–155. MR3293922 https://doi.org/10.1111/rssc.12068

WOOD, S. N., LI, Z., SHADDICK, G. and AUGUSTIN, N. H. (2017). Generalized additive models for gigadata: Modeling the U.K. Black Smoke Network daily data. *J. Amer. Statist. Assoc.* **112** 1199–1210. MR3735370 https://doi.org/10.1080/01621459.2016.1195744

ZENK, S. N., SCHULZ, A. J., MATTHEWS, S. A., ODOMS-YOUNG, A., WILBUR, J., WEGRZYN, L., GIBBS, K., BRAUNSCHWEIG, C. and STOKES, C. (2011). Activity space environment and dietary and physical activity behaviors: A pilot study. *Health Place* **17** 1150–1161.

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 816–842
https://doi.org/10.1214/21-AOAS1521

BATCH-SEQUENTIAL DESIGN AND HETEROSKEDASTIC SURROGATE MODELING FOR DELTA SMELT CONSERVATION

BY BOYA ZHANG[1,a], ROBERT B. GRAMACY[1,b], LEAH R. JOHNSON[1,c], KENNETH A. ROSE[2,e] AND ERIC SMITH[1,d]

[1]*Department of Statistics, Virginia Tech,* [a]*boya66@vt.edu,* [b]*rbg@vt.edu,* [c]*lrjohn@vt.edu,* [d]*epsmith@vt.edu*
[2]*University of Maryland Center for Environmental Science, Horn Point Laboratory,* [e]*krose@umces.edu*

Delta smelt is an endangered fish species in the San Francisco estuary that have shown an overall population decline over the past 30 years. Researchers have developed a stochastic, agent-based simulator to virtualize the system with the goal of understanding the relative contribution of natural and anthropogenic factors that might play a role in their decline. However, the input configuration space is high dimensional, running the simulator is time-consuming, and its noisy outputs change nonlinearly in both mean and variance. Getting enough runs to effectively learn input–output dynamics requires both a nimble modeling strategy and parallel evaluation. Recent advances in heteroskedastic Gaussian process (HetGP) surrogate modeling helps, but little is known about how to appropriately plan experiments for highly distributed simulation. We propose a batch sequential design scheme, generalizing one-at-a-time variance-based active learning for HetGP, as a means of keeping multicore cluster nodes fully engaged with runs. Our acquisition strategy is carefully engineered to favor selection of replicates which boost statistical and computational efficiency when training surrogates to isolate signal from noise. Design and modeling are illustrated on a range of toy examples before embarking on a large-scale smelt simulation campaign and downstream high-fidelity input sensitivity analysis.

REFERENCES

ANKENMAN, B., NELSON, B. L. and STAUM, J. (2010). Stochastic kriging for simulation metamodeling. *Oper. Res.* **58** 371–382. MR2674803 https://doi.org/10.1287/opre.1090.0754

BAKER, E., BARBILLON, P., FADIKAR, A., GRAMACY, R. B., HERBEI, R., HIGDON, D., HUANG, J., JOHNSON, L. R., MA, P. et al. (2020). Stochastic simulators: An overview with opportunities.

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

BARNETT, S. (1979). *Matrix Methods for Engineers and Scientists*. McGraw-Hill.

BAXTER, R., BROWN, L. R., CASTILLO, G., CONRAD, L., CULBERSON, S. D., DEKAR, M. P., DEKAR, M., FEYRER, F., HUNT, T. et al. (2015). An updated conceptual model of Delta Smelt biology: Our evolving understanding of an estuarine fish. Technical Report, Interagency Ecological Program, California Department of Water Resources.

BENGTSSON, H. (2018). `R.matlab`: Read and write MAT files and call MATLAB from within R. R package version 3.6.2.

BINOIS, M., GRAMACY, R. B. and LUDKOVSKI, M. (2018). Practical heteroscedastic Gaussian process modeling for large simulation experiments. *J. Comput. Graph. Statist.* **27** 808–821. MR3890872 https://doi.org/10.1080/10618600.2018.1458625

BINOIS, M., HUANG, J., GRAMACY, R. B. and LUDKOVSKI, M. (2019). Replication or exploration? Sequential design for stochastic simulation experiments. *Technometrics* **61** 7–23. MR3933655 https://doi.org/10.1080/00401706.2018.1469433

BISSET, K. R., CHEN, J., FENG, X., KUMAR, V. A. and MARATHE, M. V. (2009). EpiFast: A fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *Proceedings of the* 23*rd International Conference on Supercomputing* 430–439.

Key words and phrases. Gaussian process surrogate modeling, agent-based model, active learning, input-dependent noise, replication, sensitivity analysis.

BYRD, R. H., LU, P., NOCEDAL, J. and ZHU, C. Y. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **16** 1190–1208. MR1346301 https://doi.org/10.1137/0916069

CARNELL, R. (2020). `lhs`: Latin hypercube samples. R package version 1.0.2.

CHEN, J., MAK, S., JOSEPH, V. R. and ZHANG, C. (2019). Adaptive design for Gaussian process regression under censoring. arXiv preprint arXiv:1910.05452.

CHEVALIER, C. (2013). Fast uncertainty reduction strategies relying on Gaussian process models. Ph.D. thesis, Univ. Bern.

COLE, D. A., CHRISTIANSON, R. B. and GRAMACY, R. B. (2021). Locally induced Gaussian processes for large-scale simulation experiments. *Stat. Comput.* **31** Paper No. 33, 21. MR4244945 https://doi.org/10.1007/s11222-021-10007-9

DUAN, W., ANKENMAN, B. E., SANCHEZ, S. M. and SANCHEZ, P. J. (2017). Sliced full factorial-based Latin hypercube designs as a framework for a batch sequential design algorithm. *Technometrics* **59** 11–22. MR3604185 https://doi.org/10.1080/00401706.2015.1108233

ERICKSON, C. B., ANKENMAN, B. E., PLUMLEE, M. and SANCHEZ, S. M. (2018). Gradient based criteria for sequential design. In 2018 *Winter Simulation Conference* (*WSC*) 467–478.

FADIKAR, A., HIGDON, D., CHEN, J., LEWIS, B., VENKATRAMANAN, S. and MARATHE, M. (2018). Calibrating a stochastic, agent-based model using quantile-based emulation. *SIAM/ASA J. Uncertain. Quantificat.* **6** 1685–1706. MR3890793 https://doi.org/10.1137/17M1161233

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

GINSBOURGER, D. and LE RICHE, R. (2010). Towards Gaussian process-based optimization with finite time horizon. In *MODa* 9*–Advances in Model-Oriented Design and Analysis* 89–96. Springer.

GINSBOURGER, D., LE RICHE, R. and CARRARO, L. (2010). Kriging is well-suited to parallelize optimization. In *Computational Intelligence in Expensive Optimization Problems* 131–162. Springer.

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

GRAMACY, R. B. (2007). tgp: An R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian process models. *J. Stat. Softw.* **19** 1–46.

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

GRAMACY, R. B. and LEE, H. K. H. (2009). Adaptive design and analysis of supercomputer experiments. *Technometrics* **51** 130–145. MR2668170 https://doi.org/10.1198/TECH.2009.0015

GRAMACY, R. B. and POLSON, N. G. (2011). Particle learning of Gaussian process models for sequential design and optimization. *J. Comput. Graph. Statist.* **20** 102–118. Supplementary material available online. MR2816540 https://doi.org/10.1198/jcgs.2010.09171

GRAMACY, R. B. and TADDY, M. (2010). Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an R package for treed Gaussian process models. *J. Stat. Softw.* **33** 1–48.

HAMILTON, S. and MURPHY, D. (2018). Analysis of limiting factors across the life cycle of delta smelt (Hypomesus transpacificus). *Environ. Manag.* **62**. https://doi.org/10.1007/s00267-018-1014-9

HONG, L. J. and NELSON, B. L. (2006). Discrete optimization via simulation using COMPASS. *Oper. Res.* **54** 115–129.

JOHNSON, L. R. (2008). Microcolony and biofilm formation as a survival strategy for bacteria. *J. Theoret. Biol.* **251** 24–34. MR2945045 https://doi.org/10.1016/j.jtbi.2007.10.039

JOHNSON, M. E., MOORE, L. M. and YLVISAKER, D. (1990). Minimax and maximin distance designs. *J. Statist. Plann. Inference* **26** 131–148. MR1079258 https://doi.org/10.1016/0378-3758(90)90122-B

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

KENNEDY, J. and EBERHART, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95—International Conference on Neural Networks* **4** 1942–1948.

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

KIMMERER, W. and ROSE, K. (2018). Individual-based modeling of delta smelt population dynamics in the upper San Francisco estuary III. Effects of entrainment mortality and changes in prey. *Trans. Am. Fish. Soc.* **147** 223–243. https://doi.org/10.1002/tafs.10015

LEATHERMAN, E. R., SANTNER, T. J. and DEAN, A. M. (2018). Computer experiment designs for accurate prediction. *Stat. Comput.* **28** 739–751. MR3766041 https://doi.org/10.1007/s11222-017-9760-8

LI, Y. and DENG, X. (2021). An efficient algorithm for elastic I-optimal design of generalized linear models. *Canad. J. Statist.* **49** 438–470. MR4267928 https://doi.org/10.1002/cjs.11571

LOEPPKY, J. L., MOORE, L. M. and WILLIAMS, B. J. (2010). Batch sequential designs for computer experiments. *J. Statist. Plann. Inference* **140** 1452–1464. MR2592224 https://doi.org/10.1016/j.jspi.2009.12.004

LUND, J., HANAK, E., FLEENOR, W., BENNETT, W. and HOWITT, R. (2010). *Comparing Futures for the Sacramento, San Joaquin Delta* **3**. Univ of California Press.

LYU, X., BINOIS, M. and LUDKOVSKI, M. (2021). Evaluating Gaussian process metamodels and sequential designs for noisy level set estimation. *Stat. Comput.* **31** Paper No. 43, 21. MR4266176 https://doi.org/10.1007/s11222-021-10014-w

MACNALLY, R., THOMSON, J., KIMMERER, W., FEYRER, F., NEWMAN, K., SIH, A., BENNETT, W., BROWN, L., FLEISHMAN, E. et al. (2010). Analysis of pelagic species decline in the upper San Francisco estuary using multivariate autoregressive modeling (MAR). *Ecol. Appl.* **20** 1417–30. https://doi.org/10.1890/09-1724.1

MARREL, A., IOOSS, B., LAURENT, B. and ROUSTANT, O. (2009). Calculations of Sobol indices for the Gaussian process metamodel. *Reliab. Eng. Syst. Saf.* **94** 742–751.

MAUNDER, M. and DERISO, R. (2011). A state-space multistage life cycle model to evaluate population impacts in the presence of density dependence: Illustrated with application to delta smelt (hyposmesus transpacificus). *Can. J. Fish. Aquat. Sci.* **68** 1285–1306. https://doi.org/10.1139/f2011-071

MCKAY, M. D., BECKMAN, R. J. and CONOVER, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21** 239–245. MR0533252 https://doi.org/10.2307/1268522

MCKEAGUE, I. W., NICHOLLS, G., SPEER, K. and HERBEI, R. (2005). Statistical inversion of South Atlantic circulation in an abyssal neutral density layer. *J. Mar. Res.* **63** 683–704. https://doi.org/10.1357/0022240054663240

MEBANE, W. and SEKHON, J. (2011). Genetic optimization using derivatives: The rgenoud package for R. *J. Stat. Softw.* **42** 1–26.

MILLER, W. J., MANLY, B. F., MURPHY, D. D., FULLERTON, D. and RAMEY, R. R. (2012). An investigation of factors affecting the decline of delta smelt (Hypomesus transpacificus) in the Sacramento–San Joaquin estuary. *Reviews Fish. Sci.* **20** 1–19.

MORRIS, M. D. and MITCHELL, T. J. (1995). Exploratory designs for computational experiments. *J. Statist. Plann. Inference* **43** 381–402.

MOYLE, P. B., BROWN, L. R., DURAND, J. R. and HOBBS, J. A. (2016). Delta smelt: Life history and decline of a once-abundant species in the San Francisco estuary. *San Francisco Estuary and Watershed Science* **14**.

OAKLEY, J. E. and O'HAGAN, A. (2004). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 751–769. MR2088780 https://doi.org/10.1111/j.1467-9868.2004.05304.x

ROSE, K. A., KIMMERER, W. J., EDWARDS, K. P. and BENNETT, W. A. (2013a). Individual-based modeling of delta smelt population dynamics in the upper San Francisco estuary: I. model description and baseline results. *Trans. Am. Fish. Soc.* **142** 1238–1259. https://doi.org/10.1080/00028487.2013.799518

ROSE, K. A., KIMMERER, W. J., EDWARDS, K. P. and BENNETT, W. A. (2013b). Individual-based modeling of delta smelt population dynamics in the upper San Francisco estuary: II. Alternative baselines and good versus bad years. *Trans. Am. Fish. Soc.* **142** 1260–1272. https://doi.org/10.1080/00028487.2013.799519

RUTTER, C. M., OZIK, J., DEYOREO, M. and COLLIER, N. (2019). Microsimulation model calibration using incremental mixture approximate Bayesian computation. *Ann. Appl. Stat.* **13** 2189–2212. MR4037427 https://doi.org/10.1214/19-aoas1279

SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. With comments and a rejoinder by the authors. MR1041765

SALTELLI, A., CHAN, K. and SCOTT, E. M., eds. (2000). *Sensitivity Analysis. Wiley Series in Probability and Statistics*. Wiley, Chichester. MR1886391

SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2018). *The Design and Analysis of Computer Experiments. Springer Series in Statistics*. Springer, New York. Second edition of [MR2160708]. MR3887662

SEO, S., WALLAT, M., GRAEPEL, T. and OBERMAYER, K. (2000). Gaussian process regression: Active data selection and test point rejection. In *Proceedings of the International Joint Conference on Neural Networks* **III** 241–246. IEEE.

STEIN, M. (2012). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media, New York, NY. MR1697409 https://doi.org/10.1007/978-1-4612-1494-6

STOMPE, D. K., MOYLE, P. B., KRUGER, A. and DURAND, J. R. (2020). Comparing and integrating fish surveys in the San Francisco estuary: Why diverse long-term monitoring programs are important. *San Francisco Estuary and Watershed Science* **18**.

TADDY, M. A., LEE, H. K. H., GRAY, G. A. and GRIFFIN, J. D. (2009). Bayesian guided pattern search for robust local optimization. *Technometrics* **51** 389–401. MR2756475 https://doi.org/10.1198/TECH.2009.08007

THOMSON, J., KIMMERER, W., BROWN, L., NEWMAN, K., MAC NALLY, R., BENNETT, W., FEYRER, F. and FLEISHMAN, E. (2010). Bayesian change point analysis of abundance trends for pelagic fishes in the upper San Francisco estuary. *Ecol. Appl.* **20** 1431–48. https://doi.org/10.1890/09-0998.1

WYCOFF, N., BINOIS, M. and WILD, S. M. (2019). Sequential learning of active subspaces.

XIE, J., FRAZIER, P. I., SANKARAN, S., MARSDEN, A. and ELMOHAMED, S. (2012). Optimization of computationally expensive simulations with Gaussian processes and parameter uncertainty: Application to cardiovascular surgery. In 2012 50*th Annual Allerton Conference on Communication*, *Control*, *and Computing* (*Allerton*) 406–413. https://doi.org/10.1109/Allerton.2012.6483247

YU, H. (2002). `Rmpi`: Parallel statistical computing in R. *R News* **2** 10–14.

ZHANG, B., GRAMACY, R. B., JOHNSON, L., ROSE, K. A. and SMITH, E. (2022). Supplement to "Batch-sequential design and heteroskedastic surrogate modeling for delta smelt conservation." https://doi.org/10.1214/21-AOAS1521SUPP

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 843–865
https://doi.org/10.1214/21-AOAS1522

INTENSITY ESTIMATION ON GEOMETRIC NETWORKS WITH PENALIZED SPLINES

BY MARC SCHNEBLE[a] AND GÖRAN KAUERMANN[b]

Department of Statistics, Ludwig-Maximilians-Universität Munich, [a]*marc.schneble@stat.uni-muenchen.de,* [b]*goeran.kauermann@stat.uni-muenchen.de*

In the past decades the growing amount of network data lead to many novel statistical models. In this paper we consider so-called geometric networks. Typical examples are road networks or other infrastructure networks. Nevertheless, the neurons or the blood vessels in a human body can also be interpreted as a geometric network embedded in a three-dimensional space. A network-specific metric, rather than the Euclidean metric, is usually used in all these applications, making the analyses of network data challenging. We consider network-based point processes, and our task is to estimate the intensity (or density) of the process which allows us to detect high- and low-intensity regions of the underlying stochastic processes. Available routines that tackle this problem are commonly based on kernel smoothing methods. This paper uses penalized spline smoothing and extends this toward smooth intensity estimation on geometric networks. Furthermore, our approach easily allows incorporating covariates, enabling us to respect the network geometry in a regression model framework. Several data examples and a simulation study show that penalized spline-based intensity estimation on geometric networks is a numerically stable and efficient tool. Furthermore, it also allows estimating linear and smooth covariate effects, distinguishing our approach from already existing methodologies.

REFERENCES

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

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

BADDELEY, A., NAIR, G., RAKSHIT, S., MCSWIGGAN, G. and DAVIES, T. M. (2021). Analysing point patterns on networks—a review. *Spat. Stat.* **42** Paper No. 100435, 35. MR4233256 https://doi.org/10.1016/j.spasta.2020.100435

BARR, C. D. and SCHOENBERG, F. P. (2010). On the Voronoi estimator for the intensity of an inhomogeneous planar Poisson process. *Biometrika* **97** 977–984. MR2746166 https://doi.org/10.1093/biomet/asq047

BASSETT, R. and SHARPNACK, J. (2019). Fused density estimation: Theory and methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 839–860. MR4025399

BEER, G. (2013). The structure of extended real-valued metric spaces. *Set-Valued Var. Anal.* **21** 591–602. MR3134449 https://doi.org/10.1007/s11228-013-0255-2

BORRUSO, G. (2008). Network density estimation: A GIS approach for analysing point patterns in a network space. *Trans. GIS* **12** 377–402.

CHAN, T. M. (2012). All-pairs shortest paths for unweighted undirected graphs in $o(mn)$ time. *ACM Trans. Algorithms* **8** Art. 34, 17. MR2981912 https://doi.org/10.1145/2344422.2344424

CURRIE, I. D., DURBAN, M. and EILERS, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 259–280. MR2188985 https://doi.org/10.1111/j.1467-9868.2006.00543.x

DE BOOR, C. (1972). On calculating with B-splines. *J. Approx. Theory* **6** 50–62. MR0338617 https://doi.org/10.1016/0021-9045(72)90080-9

Key words and phrases. Intensity estimation of stochastic point processes, generalized additive models, geometric networks, penalized splines, poisson regression with offset, `spatstat` `package`.

DIGGLE, P. (1985). A kernel method for smoothing point process data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **34** 138–147.

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

EILERS, P. H. C., MARX, B. D. and DURBÁN, M. (2015). Twenty years of P-splines. *SORT* **39** 149–186. MR3467488

FAHRMEIR, L., KNEIB, T., LANG, S. and MARX, B. (2013). *Regression*: *Models*, *Methods and Applications*. Springer, Heidelberg. MR3075546 https://doi.org/10.1007/978-3-642-34333-9

GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E., AIROLDI, E. M. et al. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.

HEUSER, H. (2006). *Lehrbuch der Analysis* 1, 2. Teubner, Stuttgart. MR0618121

KAUERMANN, G. (2002). On a small sample adjustment for the profile score function in semiparametric smoothing models. *J. Multivariate Anal.* **82** 471–485. MR1921639 https://doi.org/10.1006/jmva.2001.2032

KAUERMANN, G. and OPSOMER, J. D. (2011). Data-driven selection of the spline dimension in penalized spline regression. *Biometrika* **98** 225–230. MR2804222 https://doi.org/10.1093/biomet/asq081

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

MCSWIGGAN, G. (2019). Spatial point process methods for linear networks with applications to road accident analysis. Doctoral Thesis, Univ. Western Australia.

MCSWIGGAN, G., BADDELEY, A. and NAIR, G. (2017). Kernel density estimation on a linear network. *Scand. J. Stat.* **44** 324–345. MR3658517 https://doi.org/10.1111/sjos.12255

MORADI, M. M., RODRÍGUEZ-CORTÉS, F. J. and MATEU, J. (2018). On kernel-based intensity estimation of spatial point patterns on linear networks. *J. Comput. Graph. Statist.* **27** 302–311. MR3816266 https://doi.org/10.1080/10618600.2017.1360782

MORADI, M. M., CRONIE, O., RUBAK, E., LACHIEZE-REY, R., MATEU, J. and BADDELEY, A. (2019). Resample-smoothing of Voronoi intensity estimators. *Stat. Comput.* **29** 995–1010. MR3994614 https://doi.org/10.1007/s11222-018-09850-0

O'DONNELL, D., RUSHWORTH, A., BOWMAN, A. W., SCOTT, E. M. and HALLARD, M. (2014). Flexible regression models over river networks. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 47–63. MR3148268 https://doi.org/10.1111/rssc.12024

OKABE, A., SATOH, T. and SUGIHARA, K. (2009). A kernel density estimation method for networks, its computational method and a GIS-based tool. *Int. J. Geogr. Inf. Sci.* **23** 7–32.

OKABE, A. and YAMADA, I. (2001). The K-function method on a network and its computational implementation. *Geogr. Anal.* **33** 271–290.

OKABE, A., YOMONO, H. and KITAMURA, M. (1995). Statistical analysis of the distribution of points on a network. *Geogr. Anal.* **27** 152–175.

R CORE TEAM (2013). *R*: *A Language and Environment for Statistical Computing*.

RAKSHIT, S., DAVIES, T., MORADI, M. M., MCSWIGGAN, G., NAIR, G., MATEU, J. and BADDELEY, A. (2019). Fast kernel smoothing of point patterns on a large network using two-dimensional convolution. *Int. Stat. Rev.* **87** 531–556. MR4043351 https://doi.org/10.1111/insr.12327

RASMUSSEN, J. G. and CHRISTENSEN, H. S. (2021). Point processes on directed linear networks. *Methodol. Comput. Appl. Probab.* **23** 647–667. MR4272635 https://doi.org/10.1007/s11009-020-09777-y

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

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2009). Semiparametric regression during 2003–2007. *Electron. J. Stat.* **3** 1193–1256. MR2566186 https://doi.org/10.1214/09-EJS525

RUSHWORTH, A. M., PETERSON, E. E., VER HOEF, J. M. and BOWMAN, A. W. (2015). Validation and comparison of geostatistical and spline models for spatial stream networks. *Environmetrics* **26** 327–338. MR3366967 https://doi.org/10.1002/env.2340

SCHELLHASE, C. and KAUERMANN, G. (2012). Density estimation and comparison with a penalized mixture approach. *Comput. Statist.* **27** 757–777. MR3041856 https://doi.org/10.1007/s00180-011-0289-6

SCHNEBLE, M. and KAUERMANN, G. (2022). Supplement to "Intensity estimation on geometric networks with penalized splines." https://doi.org/10.1214/21-AOAS1522SUPP

SCOTT, D. W. (2015). *Multivariate Density Estimation*: *Theory*, *Practice*, *and Visualization*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3329609

SNIJDERS, T. A. B. (1996). Stochastic actor-oriented models for network change. *J. Math. Sociol.* **21** 149–172.

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

WOOD, S. N. (2017). *Generalized Additive Models*: *An Introduction with R*. Chapman and Hall/CRC. MR2206355

WOOD, S. N. and FASIOLO, M. (2017). A generalized Fellner–Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics* **73** 1071–1081. MR3744521 https://doi.org/10.1111/biom.12666

XIE, Z. and YAN, J. (2008). Kernel density estimation of traffic accidents in a network space. *Comput*. *Environ*. *Urban Syst*. **32** 396–406.

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 866–886
https://doi.org/10.1214/21-AOAS1523

SPARSE BLOCK SIGNAL DETECTION AND IDENTIFICATION FOR SHARED CROSS-TRAIT ASSOCIATION ANALYSIS

BY JIANQIAO WANG[1,a], WANJIE WANG[2,c] AND HONGZHE LI[1,b]

[1]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania,* [a]*wangjq@upenn.edu,* [b]*hongzhe@upenn.edu*

[2]*Department of Statistics and Applied Probability, National University of Singapore,* [c]*staww@nus.edu.sg*

Genome-wide association studies (GWAS) have identified thousands of single nucleotide polymorphisms (SNPs) that are associated with complex traits. GWAS data allows us to investigate the shared genetic etiologies among different traits. However, linkage disequilibrium (LD) between the SNPs complicates the detection and identification of shared genetic effects. In this paper we model the LD by dividing the genome into LD blocks and linking the genetic variants within a block to a possible latent causal variant. An eigenvector-projected score statistic that leverages the set of variants in LD and a maxtype test statistic (Max-block) are proposed to detect the existence of cross-trait genetic association. The Max-block is easy to calculate and is shown to control the genome-wide error rate. After the detection a stepwise procedure is proposed to identify the significant blocks that explain the genetic sharing between two traits. Simulation experiments show that Max-block is more powerful than standard approaches in the sparse settings and is robust to different signal strengths or levels of sparsity. The method is applied to study shared cross-trait associations in 10 pediatric autoimmune diseases and identified several regions that explain the genetic sharing between juvenile idiopathic arthritis (JIA) and ulcerative colitis (UC) and between UC and Crohn's disease (CD). In addition, our analysis also indicates the genetic sharing in the MHC region among systemic lupus (SLE), celiac disease (CEL) and common variable immunodeficiency (CVID). Results from real data and simulation studies show that Max-block provides an important alternative to commonly used genetic correlation estimation in understanding genetic correlation among complex diseases.

REFERENCES

BEGOVICH, A. B., CHANG, M., CAILLIER, S., LEW, D., CATANESE, J. J., WANG, J., HAUSER, S. L. and OKSENBERG, J. R. (2007). The autoimmune disease–associated IL12B and IL23R polymorphisms in multiple sclerosis. *Hum. Immunol.* **68** 934–937.

BULIK-SULLIVAN, B., FINUCANE, H. K., ANTTILA, V., GUSEV, A., DAY, F. R., LOH, P.-R., DUNCAN, L., PERRY, J. R. B., PATTERSON, N., ROBINSON, E. B. et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47** 1236.

CHANG, C. C., CHOW, C. C., TELLIER, L. C. A. M., VATTIKUTI, S., PURCELL, S. M. and LEE, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4** s13742–015.

COTSAPAS, C., VOIGHT, B. F., ROSSIN, E., LAGE, K., NEALE, B. M., WALLACE, C., ABECASIS, G. R., BARRETT, J. C., BEHRENS, T. et al. (2011). Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* **7** e1002254. https://doi.org/10.1371/journal.pgen.1002254

DEMA, B., FERNÁNDEZ-ARQUERO, M., MALUENDA, C., POLANCO, I., FIGUEREDO, M. Á., EMILIO, G., URCELAY, E. and NÚÑEZ, C. (2009). Lack of association of NKX2-3, IRGM, and ATG16L1 inflammatory bowel disease susceptibility variants with celiac disease. *Hum. Immunol.* **70** 946–949.

DENNY, J. C., BASTARACHE, L., RITCHIE, M. D., CARROLL, R. J., ZINK, R., MOSLEY, J. D., FIELD, J. R., PULLEY, J. M., RAMIREZ, A. H. et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31** 1102–1110. https://doi.org/10.1038/nbt.2749

Key words and phrases. Local dependency, pleiotropy, principal component analysis, sparse simultaneous signal.

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

GAO, B., YANG, C. and LIU, J. (2021). Accurate genetic and environmental covariance estimation with composite likelihood in genome-wide association studies. *PLoS Genet.* **17** e1009293.

GIAMBARTOLOMEI, C., VUKCEVIC, D., SCHADT, E. E., FRANKE, L., HINGORANI, A. D., WALLACE, C. and PLAGNOL, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10** e1004383. https://doi.org/10.1371/journal.pgen.1004383

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

HE, X., FULLER, C. K., SONG, Y., MENG, Q., ZHANG, B. and YANG, X. (2013). Sherlock: Detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am. J. Hum. Genet.* **92** 667–680.

HILL, W. G. and ROBERTSON, A. (1968). Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38** 226–231.

HORMOZDIARI, F., VAN DE BUNT, M. and SEGRÈ (2016). Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99** 1245–1260.

LEE, HONG, Y., JIAN, G. M. E. and SANG (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28** 2540–2542.

LI, Y. R., LI, J., ZHAO, S. D. et al. (2015). Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nat. Med.* **21** 1018–1027.

LIN, C. Y., XING, G. and XING, C. (2012). Measuring linkage disequilibrium by the partial correlation coefficient. *Heredity* **109** 401–402.

LIU, Z. and LIN, X. (2019). A geometric perspective on the power of principal component association tests in multiple phenotype studies. *J. Amer. Statist. Assoc.* **114** 975–990. MR4011752 https://doi.org/10.1080/01621459.2018.1513363

LU, X., TANG, L., LI, K., ZHENG, J., ZHAO, P. and TAO, Y. (2014). Contribution of NKX2-3 polymorphisms to inflammatory bowel diseases: A meta-analysis of 35358 subjects. *Sci. Rep.* **4** 1–9.

MEGLIO, P. D., CESARE, A. D., LAGGNER, U., CHU, C.-C., NAPOLITANO, L., VILLANOVA, F., TOSI, I., CAPON, F., TREMBATH, R. C. et al. (2011). The IL23R R381Q gene variant protects against immune-mediated diseases by impairing IL-23-induced Th17 effector response in humans. *PLoS ONE* **6** e17160. https://doi.org/10.1371/journal.pone.0017160

NI, G., MOSER, G., RIPKE, S. and NEALE, B. M. (2018). Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. *Am. J. Hum. Genet.* **102** 1185–1194.

CROSS-DISORDER GROUP OF THE PSYCHIATRIC GENOMICS CONSORTIUM AND OTHERS (2013). Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis. *Lancet* **381** 1371–1379.

PASANIUC, B. and PRICE, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18** 117–127. https://doi.org/10.1038/nrg.2016.142

RESTREPO, N. A., BUTKIEWICZ, M., MCGRATH, J. A. and CRAWFORD, D. C. (2016). Shared genetic etiology of autoimmune diseases in patients from a biorepository linked to de-identified electronic health records. *Front. Genet.* **7** 185. https://doi.org/10.3389/fgene.2016.00185

SOLOVIEFF, N., COTSAPAS, C., LEE, P. H., PURCELL, S. M. and SMOLLER, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nat. Rev. Genet.* **14** 483–495. https://doi.org/10.1038/nrg3461

THE 1000 GENOMES PROJECT CONSORTIUM (2015). A global reference for human genetic variation. *Nature* **526** 68–74. https://doi.org/10.1038/nature15393

TULLIUS, S. G., BIEFER, H. R. C., LI, S., TRACHTENBERG, A. J., EDTINGER, K., QUANTE, M., KRENZIEN, F., UEHARA, H., YANG, X. et al. (2014). NAD+ protects against EAE by regulating CD4+ T-cell differentiation. *Nat. Commun.* **5** 5101. https://doi.org/10.1038/ncomms6101

WANG, J., WANG, W. and LI, H. (2022). Supplement to "Sparse block signal detection and identification for shared cross-trait association analysis." https://doi.org/10.1214/21-AOAS1523SUPP

YANG, J., ZENG, J. and GODDARD (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* **49** 1304.

YANG, J., LEE, S. H., GODDARD, M. E. and VISSCHER, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88** 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

ZHAO, S. D., CAI, T. T., CAPPOLA, T. P., MARGULIES, K. B. and LI, H. (2017). Sparse simultaneous signal detection for identifying genetically controlled disease genes. *J. Amer. Statist. Assoc.* **112** 1032–1046. MR3735358 https://doi.org/10.1080/01621459.2016.1270825

ZHU, X., FENG, T., TAYO, B. O., LIANG, J., YOUNG, J. H., FRANCESCHINI, N., SMITH, J. A., YANEK, L. R., SUN, Y. V. et al. (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet.* **96** 21–36. https://doi.org/10.1016/j.ajhg.2014.11.011

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 887–904
https://doi.org/10.1214/21-AOAS1524

COMPUTATIONALLY EFFICIENT BAYESIAN UNIT-LEVEL MODELS FOR NON-GAUSSIAN DATA UNDER INFORMATIVE SAMPLING WITH APPLICATION TO ESTIMATION OF HEALTH INSURANCE COVERAGE

BY PAUL A. PARKER[1,a], SCOTT H. HOLAN[1,2,b] AND RYAN JANICKI[3,c]

[1]*Department of Statistics, University of Missouri,* [a]*paulparker@mail.missouri.edu,* [b]*holans@missouri.edu*
[2]*Office of the Associate Director for Research and Methodology, U.S. Census Bureau*
[3]*Center for Statistical Research and Methodology, U.S. Census Bureau,* [c]*ryan.janicki@census.gov*

Statistical estimates from survey samples have traditionally been obtained via design-based estimators. In many cases these estimators tend to work well for quantities, such as population totals or means, but can fall short as sample sizes become small. In today's "information age," there is a strong demand for more granular estimates. To meet this demand, using a Bayesian pseudolikelihood, we propose a computationally efficient unit-level modeling approach for non-Gaussian data collected under informative sampling designs. Specifically, we focus on binary and multinomial data. Our approach is both multivariate and multiscale, incorporating spatial dependence at the area level. We illustrate our approach through an empirical simulation study and through a motivating application to health insurance estimates, using the American Community Survey.

REFERENCES

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

BATTESE, G. E., HARTER, R. M. and FULLER, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.* **83** 28–36.

BAUDER, M., LUERY, D. and SZELEPKA, S. (2018). Small area estimation of health insurance coverage in 2010–2016. Technical Report. Small Area Methods Branch, Social, Economic, and Housing Statistics Division, U. S. Census Bureau.

BEAL, M. J. and GHAHRAMANI, Z. (2003). The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) 453–463. Oxford Univ. Press, New York. MR2003189

BINDER, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.* **51** 279–292. MR0731144 https://doi.org/10.2307/1402588

BRADLEY, J. R., CRESSIE, N. and SHI, T. (2016). A comparison of spatial predictors when datasets could be very large. *Stat. Surv.* **10** 100–131. MR3527662 https://doi.org/10.1214/16-SS115

BRADLEY, J. R., HOLAN, S. H. and WIKLE, C. K. (2020). Bayesian hierarchical models with conjugate full-conditional distributions for dependent data from the natural exponential family. *J. Amer. Statist. Assoc.* **115** 2037–2052. MR4189775 https://doi.org/10.1080/01621459.2019.1677471

BREWER, K. R. W., EARLY, L. J. and HANIF, M. (1984). Poisson, modified Poisson and collocated sampling. *J. Statist. Plann. Inference* **10** 15–30. MR0752450 https://doi.org/10.1016/0378-3758(84)90029-6

DURANTE, D. and RIGON, T. (2019). Conditionally conjugate mean-field variational Bayes for logistic models. *Statist. Sci.* **34** 472–485. MR4017524 https://doi.org/10.1214/19-STS712

GELMAN, A. and LITTLE, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Surv. Methodol.* **23** 127–135.

GEWEKE, J. F. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Staff Report No. 148, Federal Reserve Bank of Minneapolis.

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

Key words and phrases. Bayesian analysis, informative sampling, Pólya Gamma, pseudolikelihood, small area estimation, Small Area Health Insurance Estimates (SAHIE) Program.

HUGHES, J. and HARAN, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 139–159. MR3008275 https://doi.org/10.1111/j.1467-9868.2012.01041.x

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

LINDERMAN, S., JOHNSON, M. J. and ADAMS, R. P. (2015). Dependent multinomial models made easy: Stick-breaking with the Pólya-Gamma augmentation. In *Advances in Neural Information Processing Systems* 3456–3464.

LITTLE, R. J. (2012). Calibrated Bayes, an alternative inferential paradigm for official statistics. *J. Off. Stat.* **28** 309.

PARK, D. K., GELMAN, A. and BAFUMI, J. (2006). State-level opinions from national surveys: Poststratification using multilevel logistic regression. In *Public Opinion in State Politics* Stanford Univ. Press, Stanford, CA.

PARKER, P. A., HOLAN, S. H. and JANICKI, R. (2020). Conjugate Bayesian unit-level modelling of count data under informative sampling designs. *Stat* **9** e4267, 9. MR4104231 https://doi.org/10.1002/sta4.267

PARKER, P. A., HOLAN, S. H. and JANICKI, R. (2022). Supplement to "Computationally efficient Bayesian unit-level models for non-Gaussian data under informative sampling with application to estimation of health insurance coverage." https://doi.org/10.1214/21-AOAS1524SUPP

PARKER, P. A., JANICKI, R. and HOLAN, S. H. (2019). Unit level modeling of survey data for small area estimation under informative sampling: A comprehensive overview with extensions. Preprint. Available at arXiv:1908.10488.

PFEFFERMANN, D. and SVERCHKOV, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *J. Amer. Statist. Assoc.* **102** 1427–1439. MR2412558 https://doi.org/10.1198/016214507000001094

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

SAVITSKY, T. D. and TOTH, D. (2016). Bayesian estimation under informative sampling. *Electron. J. Stat.* **10** 1677–1708. MR3522657 https://doi.org/10.1214/16-EJS1153

SI, Y., PILLAI, N. S. and GELMAN, A. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Anal.* **10** 605–625. MR3420817 https://doi.org/10.1214/14-BA924

SKINNER, C. J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys* (C. J. Skinner, D. Holt and T. M. F. Smith, eds.) 80–84. Wiley, Chichester.

STAN DEVELOPMENT TEAM (2021). Stan modeling language users guide and reference manual, version 2.26. https://mc-stan.org.

VANDENDIJCK, Y., FAES, C., KIRBY, R. S., LAWSON, A. and HENS, N. (2016). Model-based inference for small area estimation with sampling weights. *Spat. Stat.* **18** 455–473. MR3575502 https://doi.org/10.1016/j.spasta.2016.09.004

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

WINDLE, J., POLSON, N. and SCOTT, J. (2013). BayesLogit: Bayesian logistic regression. http://cran.r-project.org/web/packages/BayesLogit/index.html. R package version 0.2-4.

ZHANG, X., HOLT, J. B., LU, H., WHEATON, A. G., FORD, E. S., GREENLUND, K. J. and CROFT, J. B. (2014). Multilevel regression and poststratification for small-area estimation of population health outcomes: A case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *Am. J. Epidemiol.* **179** 1025–1033.

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 905–935
https://doi.org/10.1214/21-AOAS1525

APPROXIMATE BAYESIAN INFERENCE FOR ANALYSIS OF SPATIOTEMPORAL FLOOD FREQUENCY DATA

BY ÁRNI V. JÓHANNESSON[1,a], STEFAN SIEGERT[2,c], RAPHAËL HUSER[3,d], HAAKON BAKKA[4,e] AND BIRGIR HRAFNKELSSON[1,b]

[1]*Department of Mathematics, Faculty of Physical Sciences, School of Engineering and Natural Sciences, University of Iceland,* [a]*avj2@hi.is,* [b]*birgirhr@hi.is*

[2]*Department of Mathematics, College of Engineering, Mathematics and Physical Sciences, University of Exeter,* [c]*s.siegert@exeter.ac.uk*

[3]*Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST),* [d]*raphael.huser@kaust.edu.sa*

[4]*Department of Mathematics, University of Oslo,* [e]*bakka@r-inla.org*

Extreme floods cause casualties and widespread damage to property and vital civil infrastructure. Predictions of extreme floods, within gauged and ungauged catchments, is crucial to mitigate these disasters. In this paper a Bayesian framework is proposed for predicting extreme floods, using the generalized extreme-value (GEV) distribution. A major methodological challenge is to find a suitable parametrization for the GEV distribution when multiple covariates and/or latent spatial effects are involved and a time trend is present. Other challenges involve balancing model complexity and parsimony, using an appropriate model selection procedure and making inference based on a reliable and computationally efficient approach. We here propose a latent Gaussian modeling framework with a novel multivariate link function designed to separate the interpretation of the parameters at the latent level and to avoid unreasonable estimates of the shape and time trend parameters. Structured additive regression models, which include catchment descriptors as covariates and spatially correlated model components, are proposed for the four parameters at the latent level. To achieve computational efficiency with large datasets and richly parametrized models, we exploit a highly accurate and fast approximate Bayesian inference approach which can also be used to efficiently select models separately for each of the four regression models at the latent level. We applied our proposed methodology to annual peak river flow data from 554 catchments across the United Kingdom. The framework performed well in terms of flood predictions for both ungauged catchments and future observations at gauged catchments. The results show that the spatial model components for the transformed location and scale parameters as well as the time trend are all important, and none of these should be ignored. Posterior estimates of the time trend parameters correspond to an average increase of about 1.5% per decade with range 0.1% to 2.8% and reveal a spatial structure across the United Kingdom. When the interest lies in estimating return levels for spatial aggregates, we further develop a novel copula-based postprocessing approach of posterior predictive samples in order to mitigate the effect of the conditional independence assumption at the data level, and we demonstrate that our approach indeed provides accurate results.

REFERENCES

ALEXANDER, L. V. and JONES, P. D. (2000). Updated precipitation series for the U.K. and discussion of recent extremes. *Atmospheric Science Letters* **1** 142–150.

ANDERSON, T. W. and DARLING, D. A. (1954). A test of goodness of fit. *J. Amer. Statist. Assoc.* **49** 765–769. MR0069459

Key words and phrases. Approximate Bayesian inference, flood frequency analysis, latent Gaussian model, Max-and-Smooth, multivariate link function, spatiotemporal extremes.

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

BAKKA, H., RUE, H., FUGLSTAD, G.-A., RIEBLER, A., BOLIN, D., ILLIAN, J., KRAINSKI, E., SIMPSON, D. and LINDGREN, F. (2018). Spatial modeling with R-INLA: A review. *Wiley Interdiscip. Rev.*: *Comput. Stat.* **10** e1443. MR3873676 https://doi.org/10.1002/wics.1443

BLÖSCHL, G., HALL, J., VIGLIONE, A., PERDIGÃO, R. A. P., PARAJKA, J., MERZ, B., LUN, D., ARHEIMER, B., ARONICA, G. T. et al. (2019). Changing climate both increases and decreases European river floods. *Nature* **573** 108–111.

BOPP, G. P., SHABY, B. A. and HUSER, R. (2021). A hierarchical max-infinitely divisible spatial model for extreme precipitation. *J. Amer. Statist. Assoc.* **116** 93–106. MR4227677 https://doi.org/10.1080/01621459.2020.1750414

BURN, D. H. (1990). Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resour. Res.* **26** 2257–2265.

CASSON, E. and COLES, S. (1999). Spatial regression models for extremes. *Extremes* **1** 449–468.

CLARK, M., GANGOPADHYAY, S., HAY, L., RAJAGOPALAN, B. and WILBY, R. (2004). The schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* **5** 243–262.

COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. *Springer Series in Statistics*. Springer London, Ltd., London. MR1932132 https://doi.org/10.1007/978-1-4471-3675-0

COOLEY, D. and SAIN, S. R. (2010). Spatial hierarchical modeling of precipitation extremes from a regional climate model. *J. Agric. Biol. Environ. Stat.* **15** 381–402. MR2787265 https://doi.org/10.1007/s13253-010-0023-9

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1239641 https://doi.org/10.1002/9781119115151

CUNNANE, C. and NASH, J. (1974). Bayesian estimation of frequency of hydrological events. *Mathematical Models in Hydrology* **1**.

DADSON, S., HALL, J., MURGATROYD, A., ACREMAN, M., BATES, P., BEVEN, K., HEATHWAITE, A., HOLDEN, J., HOLMAN, I. et al. (2017). A restatement of the natural science evidence concerning catchment-based 'natural' flood management in the UK. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **473** 20160706.

DALRYMPLE, T. (1960). Flood-frequency analyses, Manual of Hydrology: Part 3. Technical Report USGPO.

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

DAVISON, A., HUSER, R. and THIBAUD, E. (2019). Spatial extremes. In *Handbook of Environmental and Ecological Statistics* M. Fuentes, J. A. Hoeting and R. L. Smith, eds.) *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 711–744. CRC Press, Boca Raton, FL. MR3889918

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

DYRRDAL, A. V., LENKOSKI, A., THORARINSDOTTIR, T. L. and STORDAL, F. (2015). Bayesian hierarchical modeling of extreme hourly precipitation in Norway. *Environmetrics* **26** 89–106. MR3324904 https://doi.org/10.1002/env.2301

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

GEIRSSON, Ó. P., HRAFNKELSSON, B. and SIMPSON, D. (2015). Computationally efficient spatial modeling of annual maximum 24-h precipitation on a fine grid. *Environmetrics* **26** 339–353. MR3366968 https://doi.org/10.1002/env.2343

GEIRSSON, Ó. P., HRAFNKELSSON, B., SIMPSON, D. and SIGURDARSON, H. (2020). LGM split sampler: An efficient MCMC sampling scheme for latent Gaussian models. *Statist. Sci.* **35** 218–233. MR4106602 https://doi.org/10.1214/19-STS727

GREHYS (1996). Presentation and review of some methods for regional flood frequency analysis. *J. Hydrol.* **186** 63–84.

HOSKING, J. R. M. and WALLIS, J. R. (2005). *Regional Frequency Analysis*: *An Approach Based on L-Moments*. Cambridge Univ. Press, Cambridge.

HOSKING, J. R. M., WALLIS, J. R. and WOOD, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics* **27** 251–261. MR0797563 https://doi.org/10.2307/1269706

HRAFNKELSSON, B., MORRIS, J. S. and BALADANDAYUTHAPANI, V. (2012). Spatial modeling of annual minimum and maximum temperatures in Iceland. *Meteorol. Atmos. Phys.* **116** 43–61.

HRAFNKELSSON, B., SIEGERT, S., HUSER, R., BAKKA, H. and JÓHANNESSON, Á. V. (2021). Max-and-Smooth: A two-step approach for approximate Bayesian inference in latent Gaussian models. *Bayesian Anal.* **16** 611–638. MR4255342 https://doi.org/10.1214/20-ba1219

HUERTA, G. and SANSÓ, B. (2007). Time-varying models for extreme values. *Environ. Ecol. Stat.* **14** 285–299. MR2405331 https://doi.org/10.1007/s10651-007-0014-3

HUSER, R. and DAVISON, A. C. (2014). Space–time modelling of extreme events. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 439–461. MR3164873 https://doi.org/10.1111/rssb.12035

HUSER, R., OPITZ, T. and THIBAUD, E. (2021). Max-infinitely divisible models and inference for spatial extremes. *Scand. J. Stat.* **48** 321–348. MR4233175 https://doi.org/10.1111/sjos.12491

HUSER, R. and WADSWORTH, J. L. (2020). Advances in statistical modeling of spatial extremes. *Wiley Interdiscip. Rev.: Comput. Stat.* e1537.

JALBERT, J., FAVRE, A.-C., BÉLISLE, C. and ANGERS, J.-F. (2017). A spatiotemporal model for extreme precipitation simulated by a climate model, with an application to assessing changes in return levels over North America. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 941–962. MR3715590 https://doi.org/10.1111/rssc.12212

JÓHANNESSON, A. V., SIEGERT, S., HUSER, R., BAKKA, H. and HRAFNKELSSON, B. (2022a). Supplement to "Approximate Bayesian inference for analysis of spatiotemporal flood frequency data" https://doi.org/10.1214/21-AOAS1525SUPPA

JÓHANNESSON, A. V., SIEGERT, S., HUSER, R., BAKKA, H. and HRAFNKELSSON, B. (2022b). R code for the paper "Approximate Bayesian inference for analysis of spatio-temporal flood frequency data" https://doi.org/10.1214/21-AOAS1525SUPPB

KJELDSEN, T. R. (2010). Modelling the impact of urbanization on flood frequency relationships in the UK. *Hydrology Research* **41** 391–405.

KJELDSEN, T. R., AHN, H. and PROSDOCIMI, I. (2017). On the use of a four-parameter kappa distribution in regional frequency analysis. *Hydrol. Sci. J.* **62** 1354–1363.

KJELDSEN, T. R. and JONES, D. A. (2006). Prediction uncertainty in a median-based index flood method using L moments. *Water Resour. Res.* **42** W07414.

KJELDSEN, T. R. and JONES, D. A. (2009a). A formal statistical model for pooled analysis of extreme floods. *Hydrology Research* **40** 465–480.

KJELDSEN, T. R. and JONES, D. A. (2009b). An exploratory analysis of error components in hydrological regression modeling. *Water Resour. Res.* **45**.

KRAINSKI, E. T., GÓMEZ-RUBIO, V., BAKKA, H., LENZI, A., CASTRO-CAMILIO, D., SIMPSON, D., LINDGREN, F. and RUE, H. (2019). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. CRC Press/CRC, New York.

KUCZERA, G. (1999). Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference. *Water Resour. Res.* **35** 1551–1557.

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

LINDLEY, D. V. (1985). *Making Decisions*, 2nd ed. Wiley, London. MR0892099

MARTINS, E. S. and STEDINGER, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resour. Res.* **36** 737–744.

NATIONAL RIVER FLOW ARCHIVE (2018). NERC CEH, Wallingford.

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

PADOAN, S. A., RIBATET, M. and SISSON, S. A. (2010). Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc.* **105** 263–277. MR2757202 https://doi.org/10.1198/jasa.2009.tm08577

RIGBY, R. A. and STASINOPOULOS, D. M. (2005). Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 507–554. MR2137253 https://doi.org/10.1111/j.1467-9876.2005.00510.x

ROBSON, A. and REED, D. (1999). Flood estimation handbook. Institute of Hydrology, Wallingford.

ROSBJERG, D. and MADSEN, H. (1995). Uncertainty measures of regional flood frequency estimators. *J. Hydrol.* **167** 209–224.

ROULSTON, M. S. and SMITH, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.* **130** 1653–1660.

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

SANG, H. and GELFAND, A. E. (2009). Hierarchical modeling for extreme values observed over space and time. *Environ. Ecol. Stat.* **16** 407–426. MR2749848 https://doi.org/10.1007/s10651-007-0078-0

SANG, H. and GELFAND, A. E. (2010). Continuous spatial process models for spatial extreme values. *J. Agric. Biol. Environ. Stat.* **15** 49–65. MR2755384 https://doi.org/10.1007/s13253-009-0010-1

SCHEFZIK, R., THORARINSDOTTIR, T. L. and GNEITING, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statist. Sci.* **28** 616–640. MR3161590 https://doi.org/10.1214/13-STS443

SIMPSON, D., RUE, H., RIEBLER, A., MARTINS, T. G. and SØRBYE, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.* **32** 1–28. MR3634300 https://doi.org/10.1214/16-STS576

STEPHENS, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *J. Amer. Statist. Assoc.* **69** 730–737.

THORARINSDOTTIR, T. L., HELLTON, K. H., STEINBAKK, G. H., SCHLICHTING, L. and ENGELAND, K. (2018). Bayesian regional flood frequency analysis for large catchments. *Water Resour. Res.* **54** 6929–6947.

VETTORI, S., HUSER, R. and GENTON, M. G. (2019). Bayesian modeling of air pollution extremes using nested multivariate max-stable processes. *Biometrics* **75** 831–841. MR4012089 https://doi.org/10.1111/biom.13051

WILBY, R. L. and QUINN, N. W. (2013). Reconstructing multi-decadal variations in fluvial flood risk using atmospheric circulation patterns. *J. Hydrol.* **487** 109–121.

YOUNGMAN, B. D. (2019). Generalized additive models for exceedances of high thresholds with an application to return level estimation for U.S. wind gusts. *J. Amer. Statist. Assoc.* **114** 1865–1879. MR4047306 https://doi.org/10.1080/01621459.2018.1529596

YOUNGMAN, B. D. (2020). evgam: Generalised Additive Extreme Value Models. R package version 0.1.4.

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 936–958
https://doi.org/10.1214/21-AOAS1526

PERMUTATION TESTS UNDER A ROTATING SAMPLING PLAN WITH CLUSTERED DATA

BY JIAHUA CHEN[1,a], YUKUN LIU[2,d], CARILYN G. TAYLOR[1,b] AND JAMES V. ZIDEK[1,c]

[1]*Department of Statistics, University of British Columbia,* [a]*jhchen@stat.ubc.ca,* [b]*cgtaylor@stat.ubc.ca,* [c]*jim@stat.ubc.ca*
[2]*School of Statistics, East China Normal University,* [d]*ykliu@sfs.ecnu.edu.cn*

The distribution of lumber strength of any grade may evolve, for example, due to climate change, forest fire, changes in processing methods, and other factors. So, in North America the forest products industry monitors the evolution of their means, percentiles, or other parameters to ensure the wood products meet the industrial standard. For administrative convenience and informativeness, one may adopt a rotating sampling plan by sampling 36 mills in the initial occasion and having six of them replaced in each successive occasion for the next five occasions. The strength data on a specified number, commonly 10 pieces of lumbers from each sampled mills, are obtained. Under such rotating plans the observations on pieces from the same mill are correlated, and the observations on samples from the same mill taken on different occasions are also correlated. Ignoring these correlations may lead to invalid inference procedures. Yet accommodating a cluster structure in parametric models is difficult and entails a high level of misspecification risk. In this paper we explore symmetry in the clustered data collected via a rotating sampling plan to develop a permutation scheme for testing various hypotheses of interest. We also introduce a semiparametric density ratio model to link the distributions of the response variable over time. The combination retains the validity of the inference methods while extracting maximum information from the sampling plan. A simulation study indicates that the proposed permutation tests firmly control the type I error whether or not the data are clustered. The use of the density ratio model improves the power of the tests. We also apply the proposed tests to data from the motivating application. The proposed permutation tests effectively address many real-world issues with trust worth inference conclusions.

REFERENCES

ANDERSON, J. A. (1979). Multivariate logistic compounds. *Biometrika* **66** 17–26. MR0529143 https://doi.org/10.1093/biomet/66.1.17

ASTM (2006). Standard practice for establishing allowable properties for visually-graded dimension lumber. American Society for Testing and Materials, West Conshohocken, PA.

BERG, E., CECERE, W. and GHOSH, M. (2014). Small area estimation for county-level farmland cash rental rates. *Journal of Survey Statistics and Methodology* **2** 1–37.

CAI, S., CHEN, J. and ZIDEK, J. V. (2017). Hypothesis testing in the presence of multiple samples under density ratio models. *Statist. Sinica* **27** 761–783. MR3674695

CAI, Y., CAI, J., CHEN, J., GOLCHI, S., GUAN, M., KARIM, M. E., LIU, Y., TOMAL, J., XIONG, C. et al. (2016). An empirical experiment to assess the relationship between the tensile and bending strengths of lumber. The University of British Columbia, Department of Statistics, Technical Report # 276.

CHEN, J. and LIU, Y. (2013). Quantile and quantile-function estimations under density ratio model. *Ann. Statist.* **41** 1669–1692. MR3113825 https://doi.org/10.1214/13-AOS1129

CHEN, J., VARIYATH, A. M. and ABRAHAM, B. (2008). Adjusted empirical likelihood and its properties. *J. Comput. Graph. Statist.* **17** 426–443. MR2439967 https://doi.org/10.1198/106186008X321068

CHEN, J., LI, P., LIU, Y. and ZIDEK, J. V. (2021). Composite empirical likelihood for multisample clustered data. *J. Nonparametr. Stat.* **33** 60–81. MR4261898 https://doi.org/10.1080/10485252.2021.1914337

Key words and phrases. Density ratio model, empirical likelihood, exchangeability, multiple sample, percentila, rank test.

DATTA, S. and SATTEN, G. A. (2005). Rank-sum tests for clustered data. *J. Amer. Statist. Assoc.* **100** 908–915. MR2201018 https://doi.org/10.1198/016214504000001583

DATTA, S. and SATTEN, G. A. (2008). A signed-rank test for clustered data. *Biometrics* **64** 501–507, 667. MR2432420 https://doi.org/10.1111/j.1541-0420.2007.00923.x

FRANCISCO, C. A. and FULLER, W. A. (1991). Quantile estimation with a complex survey design. *Ann. Statist.* **19** 454–469. MR1091862 https://doi.org/10.1214/aos/1176347993

HEMERIK, J. and GOEMAN, J. (2018). Exact testing with random permutations. *TEST* **27** 811–825. MR3878362 https://doi.org/10.1007/s11749-017-0571-1

HEMERIK, J., SOLARI, A. and GOEMAN, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika* **106** 635–649. MR3992394 https://doi.org/10.1093/biomet/asz021

KARNA, J. P. and NATH, D. C. (2015). Rotationn sampling: Introduction and review of recent developments. *J. Assam Sci. Soc.* **56** 90–111.

KEZIOU, A. and LEONI-AUBIN, S. (2008). On empirical likelihood for semiparametric two-sample density ratio models. *J. Statist. Plann. Inference* **138** 915–928. MR2384498 https://doi.org/10.1016/j.jspi.2007.02.009

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

NIJMAN, T., VERBEEK, M. and VAN SOEST, A. (1991). The efficiency of rotating-panel designs in an analysis-of-variance model. *J. Econometrics* **49** 373–399. MR1129125 https://doi.org/10.1016/0304-4076(91)90003-V

OWEN, A. (2001). *Empirical Likelihood*. CRC Press/CRC, New York.

OWEN, A. B. (2013). Self-concordance for empirical likelihood. *Canad. J. Statist.* **41** 387–397. MR3101590 https://doi.org/10.1002/cjs.11183

PARK, Y. S., CHOI, J. W. and KIM, K. W. (2007). A balanced multi-level rotation sampling design and its efficient composite estimators. *J. Statist. Plann. Inference* **137** 594–610. MR2298960 https://doi.org/10.1016/j.jspi.2005.12.007

PESARIN, F. and SALMASO, L. (2010). *Permutation Tests for Complex Data*: *Theory, Applications and Software*. John Wiley & Sons.

PFEFFERMANN, D. and SVERCHKOV, M. (2009). Inference under informative sampling. In *Handbook of Statistics* **29** 455–487. Elsevier.

QIN, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* **85** 619–630. MR1665814 https://doi.org/10.1093/biomet/85.3.619

QIN, J. and ZHANG, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **84** 609–618. MR1603924 https://doi.org/10.1093/biomet/84.3.609

RAO, J. N. K. and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79** 811–822. MR1209480 https://doi.org/10.1093/biomet/79.4.811

ROSNER, B., GLYNN, R. J. and LEE, M.-L. T. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics* **62** 185–192. MR2226572 https://doi.org/10.1111/j.1541-0420.2005.00389.x

SOETAERT, K. (2009). rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations. R package 1.6.

SOETAERT, K. and HERMAN, P. M. J. (2009). *A Practical Guide to Ecological Modelling*: *Using R as a Simulation Platform*. Springer, New York. MR2492334 https://doi.org/10.1007/978-1-4020-8624-3

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

VERRILL, S., KRETSCHMANN, D. E. and EVANS, J. W. (2015). Simulations of strength property monitoring tests. Unpublished manuscript. Forest Products Laboratory, Madison, Wisconsin. Available at http://www1.fpl.fs.fed.us/monit.pdf.

ZIDEK, J. V. and LUM, C. (2018). Statistical challenges in assessing the engineering properties of forest products. *Annu. Rev. Stat. Appl.* **5** 237–267. MR3774747 https://doi.org/10.1146/annurev-statistics-041715-033633

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 959–981
https://doi.org/10.1214/21-AOAS1527

INFERENCE FOR STOCHASTIC KINETIC MODELS FROM MULTIPLE DATA SOURCES FOR JOINT ESTIMATION OF INFECTION DYNAMICS FROM AGGREGATE REPORTS AND VIROLOGICAL DATA

BY OKSANA A. CHKREBTII[1,a], YURY E. GARCÍA[2,b], MARCOS A. CAPISTRÁN[2,c] AND DANIEL E. NOYOLA[3,d]

[1]*Department of Statistics, The Ohio State University,* [a]*oksana@stat.osu.edu*
[2]*Área de Matemáticas Básicas, Centro de Investigación en Matemáticas,* [b]*yury@cimat.mx,* [c]*marcos@cimat.mx*
[3]*Department of Microbiology, Faculty of Medicine, Universidad Autónoma de San Luis Potosí,* [d]*dnoyola@uaslp.mx*

Before the current pandemic, influenza and respiratory syncytial virus (RSV) were the leading etiological agents of seasonal acute respiratory infections (ARI) around the world. In this setting, medical doctors typically based the diagnosis of ARI on patients' symptoms alone and did not routinely conduct virological tests necessary to identify individual viruses, limiting the ability to study the interaction between multiple pathogens and to make public health recommendations. We consider a stochastic kinetic model (SKM) for two interacting ARI pathogens circulating in a large population and an empirically-motivated background process for infections with other pathogens causing similar symptoms. An extended marginal sampling approach, based on the linear noise approximation to the SKM, integrates multiple data sources and additional model components. We infer the parameters defining the pathogens' dynamics and interaction within a Bayesian model and explore the posterior trajectories of infections for each illness based on aggregate infection reports from six epidemic seasons collected by the state health department and a subset of virological tests from a sentinel program at a general hospital in San Luis Potosí, México. We interpret the results and make recommendations for future data collection strategies.

REFERENCES

ÅNESTAD, G. (1982). Interference between outbreaks of respiratory syncytial virus and influenza virus infection. *Lancet* **1** 502. https://doi.org/10.1016/s0140-6736(82)91466-0

ÅNESTAD, G. (1987). Surveillance of respiratory viral infections by rapid immunofluorescence diagnosis, with emphasis on virus interference. *Epidemiol. Infect.* **99** 523–531. https://doi.org/10.1017/s0950268800068023

ÅNESTAD, G. and NORDBO, S. A. (2009). Interference between outbreaks of respiratory viruses. *Euro Surveill.* **14** 19359. https://doi.org/10.2807/ese.14.41.19359-en

ABAT, C., CHAUDET, H., ROLAIN, J.-M., COLSON, P. and RAOULT, D. (2016). Traditional and syndromic surveillance of infectious diseases and pathogens. *Int. J. Infect. Dis.* **48** 22–28. https://doi.org/10.1016/j.ijid.2016.04.021

ADAMS, B. and BOOTS, M. (2007). The influence of immune cross-reaction on phase structure in resonant solutions of a multi-strain seasonal SIR model. *J. Theoret. Biol.* **248** 202–211. MR2483194 https://doi.org/10.1016/j.jtbi.2007.04.023

ALLEN, L. J. S. (2008). An introduction to stochastic epidemic models. In *Mathematical Epidemiology*. *Lecture Notes in Math.* **1945** 81–130. Springer, Berlin. MR2428373 https://doi.org/10.1007/978-3-540-78911-6_3

AMINI, R., GILCA, R., BOUCHER, F. D., CHAREST, H. and SERRES, G. D. (2019). Respiratory syncytial virus contributes to more severe respiratory morbidity than influenza in children < 2 years during seasonal influenza peaks. *Infection* **47** 595–601. https://doi.org/10.1007/s15010-019-01287-5

ANDERSON, L. J., HIERHOLZER, J. C., TSOU, C., HENDRY, R. M., FERNIE, B. F., STONE, Y. and MCINTOSH, K. (1985). Antigenic characterization of respiratory syncytial virus strains with monoclonal antibodies. *J. Infect. Dis.* **151** 626–633. https://doi.org/10.1093/infdis/151.4.626

Key words and phrases. Stochastic kinetic models, acute respiratory disease, Bayesian modeling, linear noise approximation, influenza, RSV.

AUSTRALIAN GOVERNMENT DEPARTMENT OF HEALTH (2017). Australian influenza surveillance report. Available at http://www.health.gov.au/internet/main/publishing.nsf/Content/8FC4EA9E4C6E3F5CCA2581D4001BBC9A/File/ozflu-surveil-no12-2017.pdf.

BASHEY, F. (2015). Within-host competitive interactions as a mechanism for the maintenance of parasite diversity. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **370**. https://doi.org/10.1098/rstb.2014.0301

BHATTACHARYYA, S., GESTELAND, P. H., KORGENSKI, K., BJØRNSTAD, O. N. and ADLER, F. R. (2015). Cross-immunity between strains explains the dynamical pattern of paramyxoviruses. *Proc. Natl. Acad. Sci., India, Sect. B Biol. Sci.* **112** 13396–13400. https://doi.org/10.1073/pnas.1516698112

BLOOM-FESHBACH, K., ALONSO, W. J., CHARU, V., TAMERIUS, J., SIMONSEN, L., MILLER, M. A. and VIBOUD, C. (2013). Latitudinal variations in seasonal activity of influenza and respiratory syncytial virus (RSV): A global comparative review. *PLoS ONE* **8** e54445. https://doi.org/10.1371/journal.pone.0054445

BOYS, R. J., WILKINSON, D. J. and KIRKWOOD, T. B. L. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Stat. Comput.* **18** 125–135. MR2390814 https://doi.org/10.1007/s11222-007-9043-x

BRYNJARSDÓTTIR, J. and O'HAGAN, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Probl.* **30** 114007, 24. MR3274591 https://doi.org/10.1088/0266-5611/30/11/114007

CHAN, K. P., WONG, C. M., CHIU, S. S. S., CHAN, K. H., WANG, X. L., CHAN, E. L. Y., PEIRIS, J. S. M. and YANG, L. (2014). A robust parameter estimation method for estimating disease burden of respiratory viruses. *PLoS ONE* **9** e90126. https://doi.org/10.1371/journal.pone.0090126

CHARU, V., CHOWELL, G., MEJIA, L. S. P., ECHEVARRÍA-ZUNO, S., BORJA-ABURTO, V. H., SIMONSEN, L., MILLER, M. A. and VIBOUD, C. (2011). Mortality burden of the A/H1N1 pandemic in Mexico: A comparison of deaths and years of life lost to seasonal influenza. *Clin. Infect. Dis.* **53** 985–993. https://doi.org/10.1093/cid/cir644

CHAW, L., KAMIGAKI, T., BURMAA, A., URTNASAN, C., OD, I., NYAMAA, G., NYMADAWA, P. and OSHITANI, H. (2016). Burden of influenza and respiratory syncytial virus infection in pregnant women and infants under 6 months in Mongolia: A prospective cohort study. *PLoS ONE* **11** e0148421. https://doi.org/10.1371/journal.pone.0148421

CHKREBTII, O. A., GARCÍA, Y. E., CAPISTRÁN, M. A. and NOYOLA, D. E. (2022). Supplement to "Inference for stochastic kinetic models from multiple data sources for joint estimation of infection dynamics from aggregate reports and virological data." https://doi.org/10.1214/21-AOAS1527SUPPA, https://doi.org/10.1214/21-AOAS1527SUPPB

CHOI, B. and REMPALA, G. A. (2011). Inference for discretely observed stochastic kinetic networks with applications to epidemic modeling. *Biostatistics* **13** 153–165. https://doi.org/10.1093/biostatistics/kxr019

DOWELL, S. F., ANDERSON, L. J., GARY, H. E., ERDMAN, D. D., PLOUFFE, J. F., FILE, T. M., MARSTON, B. J. and BREIMAN, R. F. (1996). Respiratory syncytial virus is an important cause of community-acquired lower respiratory infection among hospitalized adults. *J. Infect. Dis.* **174** 456–462. https://doi.org/10.1093/infdis/174.3.456

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

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

FEARNHEAD, P., 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

FERGUSON, N. M., GALVANI, A. P. and BUSH, R. M. (2003). Ecological and immunological determinants of influenza evolution. *Nature* **422** 428–433. https://doi.org/10.1038/nature01509

FINKENSTÄDT, B., WOODCOCK, D. J., KOMOROWSKI, M., HARPER, C. V., DAVIS, J. R. E., WHITE, M. R. H. and RAND, D. A. (2013). Quantifying intrinsic and extrinsic noise in gene transcription using the linear noise approximation: An application to single cell data. *Ann. Appl. Stat.* **7** 1960–1982. MR3161709 https://doi.org/10.1214/13-AOAS669

FINTZI, J., WAKEFIELD, J. and MININ, V. N. (2020). A linear noise approximation for stochastic epidemic models fit to partially observed incidence counts. Available at arXiv:2001.05099.

FIORE, A. E., SHAY, D. K., BRODER, K., ISKANDER, J. K., UYEKI, T. M. et al. (2008). Prevention and control of influenza recommendations of the advisory committee on immunization practices. *MMWR, Recommendations and Reports: Morbidity and Morality Weekly Report. Recommendations and Reports* **57** 1–60.

FLEMING, D. M., TAYLOR, R. J., LUSTIG, R. L., SCHUCK-PAIM, C., HAGUINET, F., WEBB, D. J., LOGIE, J., MATIAS, G. and TAYLOR, S. (2015). Modelling estimates of the burden of Respiratory Syncytial virus infection in adults and the elderly in the United Kingdom. *BMC Infect. Dis.* **15** 1–12. https://doi.org/10.1186/s12879-015-1218-z

GEYER, C. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics*, *Proceedings of the* 23*rd Symposium on the Interface* **156** Amer. Statist. Assoc..

GILLESPIE, D. T. (2007). Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* **58** 35–55. https://doi.org/10.1146/annurev.physchem.58.032806.104637

GJINI, E., VALENTE, C., SÁ-LEÃO, R. and GOMES, M. G. M. (2016). How direct competition shapes coexistence and vaccine effects in multi-strain pathogen systems. *J. Theoret. Biol.* **388** 50–60. https://doi.org/10.1016/j.jtbi.2015.09.031

GOLIGHTLY, A., HENDERSON, D. A. and SHERLOCK, C. (2015). Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Stat. Comput.* **25** 1039–1055. MR3375634 https://doi.org/10.1007/s11222-014-9469-x

GOLIGHTLY, A. and WILKINSON, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus* **1** 807–820. https://doi.org/10.1098/rsfs.2011.0047

GÓMEZ-VILLA, R. J., COMAS-GARCÍA, A., LÓPEZ-ROJAS, V., PÉREZ-GONZÁLEZ, L. F., SÁNCHEZ-ALVARADO, J., SALAZAR-ZARAGOZA, R., RUIZ-GONZÁLEZ, J. M., ALPUCHE-SOLÍS, Á. G. and NOYOLA, D. E. (2008). Effect of an infection control program on the frequency of nosocomial viral respiratory infections. *Infect. Control Hosp. Epidemiol.* **29** 556–558. https://doi.org/10.1086/588000

GRIMSHAW, R. (1990). *Nonlinear Ordinary Differential Equations. Applied Mathematics and Engineering Science Texts* **2**. Blackwell Scientific Publications Ltd., Oxford. MR1098714

GRÖNDAHL, B., ANKERMANN, T., VON BISMARCK, P., ROCKAHR, S., KOWALZIK, F., GEHRING, S., MEYER, C., KNUF, M. and PUPPE, W. (2013). The 2009 pandemic influenza A(H1N1) coincides with changes in the epidemiology of other viral pathogens causing acute respiratory tract infections in children. *Infection* **42** 303–308. https://doi.org/10.1007/s15010-013-0545-5

HASHEM, M. (2003). Respiratory syncytial virus in healthy adults: The cost of a cold. *J. Clin. Virol.* **27** 14–21. https://doi.org/10.1016/s1386-6532(03)00047-7

HEY, K. L., MOMIJI, H., FEATHERSTONE, K., DAVIS, J. R. E., WHITE, M. R. H., RAND, D. A. and FINKENSTÄDT, B. (2015). A stochastic transcriptional switch model for single cell imaging data. *Biostatistics* **16** 655–669. MR3449834 https://doi.org/10.1093/biostatistics/kxv010

HUPPERT, A. and KATRIEL, G. (2013). Mathematical modelling and prediction in infectious disease epidemiology. *Clin. Microbiol. Infect.* **19** 999–1005. https://doi.org/10.1111/1469-0691.12308

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** 1–10. https://doi.org/10.1186/1471-2105-10-343

KURI-MORALES, P., GALVÁN, F., CRAVIOTO, P., ROSAS, L. A. Z. and TAPIA-CONYER, R. (2006). Mortalidad en México por influenza y neumonía (1990–2005). *Salud Pública Méx.* **48** 379–384. https://doi.org/10.1590/s0036-36342006000500004

MANGTANI, P., HAJAT, S., KOVATS, S., WILKINSON, P. and ARMSTRONG, B. (2006). The association of respiratory syncytial virus infection and influenza with emergency admissions for respiratory disease in London: An analysis of routine surveillance data. *Clin. Infect. Dis.* **42** 640–646. https://doi.org/10.1086/499810

MARTCHEVA, M., BOLKER, B. M. and HOLT, R. D. (2007). Vaccine-induced pathogen strain replacement: What are the mechanisms? *J. R. Soc. Interface* **5** 3–13. https://doi.org/10.1098/rsif.2007.0236

MASASHIKAMO and AKIRASASAKI (2002). The effect of cross-immunity and seasonal forcing in a multi-strain epidemic model. *Phys. D* **165** 228–241. https://doi.org/10.1016/s0167-2789(02)00389-5

MESKILL, S. D., REVELL, P. A., CHANDRAMOHAN, L. and CRUZ, A. T. (2017). Prevalence of co-infection between respiratory syncytial virus and influenza in children. *Am. J. Emerg. Med.* **35** 495–498. https://doi.org/10.1016/j.ajem.2016.12.001

MIDEO, N., ALIZON, S. and DAY, T. (2008). Linking within- and between-host dynamics in the evolutionary epidemiology of infectious diseases. *Trends Ecol. Evol.* **23** 511–517. https://doi.org/10.1016/j.tree.2008.05.009

MÍGUEZ, A., IFTIMI, A. and MONTES, F. (2016). Temporal association between the influenza virus and respiratory syncytial virus (RSV): RSV as a predictor of seasonal influenza. *Epidemiol. Infect.* **144** 2621–2632. https://doi.org/10.1017/s095026881600090x

MUFSON, M. A., BELSHE, R. B., ORVELL, C. and NORRBY, E. (1988). Respiratory syncytial virus epidemics: Variable dominance of subgroups A and B strains among children, 1981–1986. *J. Infect. Dis.* **157** 143–148. https://doi.org/10.1093/infdis/157.1.143

MUNYWOKI, P. K., KOECH, D. C., AGOTI, C. N., LEWA, C., CANE, P. A., MEDLEY, G. F. and NOKES, D. J. (2013). The source of respiratory syncytial virus infection in infants: A household cohort study in rural Kenya. *J. Infect. Dis.* **209** 1685–1692. https://doi.org/10.1093/infdis/jit828

NISHIMURA, N., NISHIO, H., LEE, M. J. and UEMURA, K. (2005). The clinical features of respiratory syncytial virus: Lower respiratory tract infection after upper respiratory tract infection due to influenza virus. *Pediatr. Int.* **47** 412–416. https://doi.org/10.1111/j.1442-200x.2005.02099.x

PERET, T. C., GOLUB, J. A., ANDERSON, L. J., HALL, C. B. and SCHNABEL, K. C. (1998). Circulation patterns of genetically distinct group A and B strains of human respiratory syncytial virus in a community. *J. Gen. Virol.* **79** 2221–2229. https://doi.org/10.1099/0022-1317-79-9-2221

PINKY, L. and DOBROVOLNY, H. M. (2016). Coinfections of the respiratory tract: Viral competition for resources. *PLoS ONE* **11** e0155589. https://doi.org/10.1371/journal.pone.0155589

RAMBAUT, A., PYBUS, O. G., NELSON, M. I., VIBOUD, C., TAUBENBERGER, J. K. and HOLMES, E. C. (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453** 615–619. https://doi.org/10.1038/nature06945

REICH, N. G., SHRESTHA, S., KING, A. A., ROHANI, P., LESSLER, J., KALAYANAROOJ, S., YOON, I.-K., GIBBONS, R. V., BURKE, D. S. et al. (2013). Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. *J. R. Soc. Interface* **10** 20130414. https://doi.org/10.1098/rsif.2013.0414

ROHANI, P. (1999). Opposite patterns of synchrony in sympatric disease metapopulations. *Science* **286** 968–971. https://doi.org/10.1126/science.286.5441.968

SHINJOH, M., OMOE, K., SAITO, N., MATSUO, N. and NEROME, K. (2000). In vitro growth profiles of respiratory syncytial virus in the presence of influenza virus. *Acta Virol.* **44** 91–97.

SHRESTHA, S., KING, A. A. and ROHANI, P. (2011). Statistical inference for multi-pathogen systems. *PLoS Comput. Biol.* **7** e1002135, 14. MR2845073 https://doi.org/10.1371/journal.pcbi.1002135

SHRESTHA, S., FOXMAN, B., WEINBERGER, D. M., STEINER, C., VIBOUD, C. and ROHANI, P. (2013). Identifying the interaction between influenza and pneumococcal pneumonia using incidence data. *Sci. Transl. Med.* **5** 191ra84. https://doi.org/10.1126/scitranslmed.3005982

SIETTOS, C. I. and RUSSO, L. (2013). Mathematical modeling of infectious disease dynamics. *Virulence* **4** 295–306. https://doi.org/10.4161/viru.24041

SIMONSEN, L., SPREEUWENBERG, P., LUSTIG, R., TAYLOR, R. J., FLEMING, D. M., KRONEMAN, M., KERKHOVE, M. D. V., MOUNTS, A. W. and PAGET, W. J. (2013). Global mortality estimates for the 2009 influenza pandemic from the GLaMOR project: A modeling study. *PLoS Med.* **10** e1001558. https://doi.org/10.1371/journal.pmed.1001558

STAR, L. and MOGHADAS, S. (2010). The role of mathematical modelling in public health planning and decision making. Available at https://nccid.ca/publications/the-role-of-mathematical-modelling-in-public-health-planning-and-decision-making/.

THOMAS, P., MATUSCHEK, H. and GRIMA, R. (2012). Intrinsic noise analyzer: A software package for the exploration of stochastic biochemical kinetics using the system size expansion. *PLoS ONE* **7** e38518. https://doi.org/10.1371/journal.pone.0038518

THOMPSON, W. W., SHAY, D. K., WEINTRAUB, E., BRAMMER, L., COX, N., ANDERSON, L. J. and FUKUDA, K. (2003). Mortality associated with influenza and respiratory syncytial virus in the United States. *J. Am. Med. Assoc.* **289** 179. https://doi.org/10.1001/jama.289.2.179

TOULOUPOU, P., FINKENSTÄDT, B. and SPENCER, S. E. F. (2020). Scalable Bayesian inference for coupled hidden Markov and semi-Markov models. *J. Comput. Graph. Statist.* **29** 238–249. MR4116038 https://doi.org/10.1080/10618600.2019.1654880

TROEGER, C., BLACKER, B., KHALIL, I. A., RAO, P. C., CAO, J., ZIMSEN, S. R. M., ALBERTSON, S. B., DESHPANDE, A., FARAG, T. et al. (2018). Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: A systematic analysis for the global burden of disease study 2016. *Lancet Infect. Dis.* **18** 1191–1210. https://doi.org/10.1016/s1473-3099(18)30310-4

VAN BAALEN, M. and SABELIS, M. W. (1995). The dynamics of multiple infection and the evolution of virulence. *Amer. Nat.* **146** 881–910. https://doi.org/10.1086/285830

VAN KAMPEN, N. G. (1992). *Stochastic Processes in Physics and Chemistry*. Elsevier.

VASCO, D. A., WEARING, H. J. and ROHANI, P. (2007). Tracking the dynamics of pathogen interactions: Modeling ecological and immune-mediated processes in a two-pathogen single-host system. *J. Theoret. Biol.* **245** 9–25. MR2306453 https://doi.org/10.1016/j.jtbi.2006.08.015

VELASCO-HERNÁNDEZ, J. X., NÚÑEZ-LÓPEZ, M., COMAS-GARCÍA, A., CHERPITEL, D. E. N. and OCAMPO, M. C. (2015). Superinfection between influenza and RSV alternating patterns in San Luis Potosí state, México. *PLoS ONE* **10** e0115674. https://doi.org/10.1371/journal.pone.0115674

VENTER, M., MADHI, S. A., TIEMESSEN, C. T. and SCHOUB, B. D. (2001). Genetic diversity and molecular epidemiology of respiratory syncytial virus over four consecutive seasons in South Africa: Identification of new subgroup A and B genotypes. *J. Gen. Virol.* **82** 2117–2124. https://doi.org/10.1099/0022-1317-82-9-2117

VIZCARRA-UGALDE, S., RICO-HERNÁNDEZ, M., MONJARÁS-ÁVILA, C., BERNAL-SILVA, S., GARROCHO-RANGEL, M. E., OCHOA-PÉREZ, U. R. and NOYOLA, D. E. (2016). Intensive care unit admission and death rates of infants admitted with respiratory syncytial virus lower respiratory tract infection in Mexico. *Pediatr. Infect. Dis. J.* **35** 1199–1203. https://doi.org/10.1097/INF.0000000000001262

WEST, M. and HARRISON, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1482232

WHITAKER, G. A., GOLIGHTLY, A., BOYS, R. J. and SHERLOCK, C. (2017). Bayesian inference for diffusion-driven mixed-effects models. *Bayesian Anal.* **12** 435–463. MR3620740 https://doi.org/10.1214/16-BA1009

WILKINSON, D. J. (2006). *Stochastic Modelling for Systems Biology*. *Chapman & Hall/CRC Mathematical and Computational Biology Series*. CRC Press/CRC, Boca Raton, FL. MR2222876

WILKINSON, D. J. (2011). *Stochastic Modelling for Systems Biology*, 2nd ed. *Chapman & Hall/CRC Mathematical and Computational Biology*. Taylor & Francis. MR2222876 https://doi.org/10.7287/PEERJ.PREPRINTS.1686V1

WORLD HEALTH ORGANIZATION (2017). Mortality and global health estimates. Available at http://www.who.int/gho/mortality_burden_disease/life_tables/situation_trends/en/.

WU, Z., SUN, X., CHU, Y., SUN, J., QIN, G., YANG, L., QIN, J., XIAO, Z., REN, J. et al. (2016). Coherence of influenza surveillance data across different sources and age groups, Beijing, China, 2008–2015. *PLoS ONE* **11** e0169199. https://doi.org/10.1371/journal.pone.0169199

ZAMBON, M., STOCKTON, J., CLEWLEY, J. and FLEMING, D. (2001). Contribution of influenza and respiratory syncytial virus to community cases of influenza-like illness: An observational study. *Lancet* **358** 1410–1416. https://doi.org/10.1016/s0140-6736(01)06528-x

ZHANG, X.-S., ANGELIS, D. D., WHITE, P. J., CHARLETT, A., PEBODY, R. G. and MCCAULEY, J. (2013). Co-circulation of influenza A virus strains and emergence of pandemic via reassortment: The role of cross-immunity. *Epidemics* **5** 20–33. https://doi.org/10.1016/j.epidem.2012.10.003

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 982–998
https://doi.org/10.1214/21-AOAS1528

MULTISTATE CAPTURE–RECAPTURE MODELS FOR IRREGULARLY SAMPLED DATA

BY SINA MEWS[1,a], ROLAND LANGROCK[1,b], RUTH KING[2,c] AND NICOLA QUICK[3,d]

[1]*Department of Business Administration and Economics, Bielefeld University,* [a]*sina.mews@uni-bielefeld.de,* [b]*roland.langrock@uni-bielefeld.de*

[2]*School of Mathematics, University of Edinburgh,* [c]*Ruth.King@ed.ac.uk*

[3]*School of Biology, University of St Andrews,* [d]*nicola.quick@duke.edu*

Multistate capture-recapture data comprise individual-specific sighting histories, together with information on individuals' states related, for example, to breeding status, infection level, or geographical location. Such data are often analysed using the Arnason–Schwarz model, where transitions between states are modelled using a discrete-time Markov chain, making the model most easily applicable to regular time series. When time intervals between capture occasions are not of equal length, more complex time-dependent constructions may be required, increasing the number of parameters to estimate, decreasing interpretability, and potentially leading to reduced precision. Here we develop a multi-state model based on a state process operating in continuous time, which can be regarded as an analogue of the discrete-time Arnason–Schwarz model for irregularly sampled data. Statistical inference is carried out by regarding the capture-recapture data as realisations from a continuous-time hidden Markov model, which allows the associated efficient algorithms to be used for maximum likelihood estimation and state decoding. To illustrate the feasibility of the modelling framework, we use a long-term survey of bottlenose dolphins where capture occasions are not regularly spaced through time. Here, we are particularly interested in seasonal effects on the movement rates of the dolphins along the Scottish east coast. The results reveal seasonal movement patterns between two core areas of their range, providing information that will inform conservation management.

REFERENCES

ADAMS, A. J., WOLFE, R. K., PINE, W. E. and THORNTON, B. L. (2006). Efficacy of PIT tags and an autonomous antenna system to study the juvenile life stage of an estuarine-dependent fish. *Estuaries Coast* **29** 311–317.

AMOROS, R., KING, R., TOYODA, H., KUMADA, T., JOHNSON, P. J. and BIRD, T. G. (2019). A continuous-time hidden Markov model for cancer surveillance using serum biomarkers with application to hepatocellular carcinoma. *Metron* **77** 67–86. MR3985190 https://doi.org/10.1007/s40300-019-00151-8

ARNASON, A. N. (1973). The estimation of population size, migration rates and survival in a stratified population. *Res. Popul. Ecol.* **15** 1–8.

BARBOUR, A. B., PONCIANO, J. M. and LORENZEN, K. (2013). Apparent survival estimation from continuous mark-recapture/resighting data. *Methods Ecol. Evol.* **4** 846–853.

BECKER, N. G. (1984). Estimating population size from capture-recapture experiments in continuous time. *Aust. J. Stat.* **26** 1–7. MR0746010 https://doi.org/10.1111/j.1467-842x.1984.tb01261.x

BORCHERS, D. L. and LANGROCK, R. (2015). Double-observer line transect surveys with Markov-modulated Poisson process models for animal availability. *Biometrics* **71** 1060–1069. MR3436731 https://doi.org/10.1111/biom.12341

BORCHERS, D., DISTILLER, G., FOSTER, R., HARMSEN, B. and MILAZZO, L. (2014). Continuous-time spatially explicit capture-recapture models, with an application to a jaguar camera-trap survey. *Methods Ecol. Evol.* **5** 656–66.

BROWNIE, C., HINES, J. E., NICHOLS, J. D., POLLOCK, K. H. and HESTBECK, J. B. (1993). Capture-recapture studies for multiple strata including non-Markovian transitions. *Biometrics* **49** 1173–1187.

Key words and phrases. Arnason–Schwarz model, continuous-time Markov chain, hidden Markov model, maximum likelihood.

BUCKLAND, S. T., NEWMAN, K. B., THOMAS, L. and KOESTERS, N. B. (2004). State-space models for the dynamics of wild animal populations. *Ecol. Model.* **171** 157–175.

CAMERON, C., BARKER, R., FLETCHER, D., SLOOTEN, E. and DAWSON, S. (1999). Modelling survival of Hector's dolphins around Banks Peninsula, New Zealand. *J. Agric. Biol. Environ. Stat.* **4** 126–135. MR1812078 https://doi.org/10.2307/1400593

CHAO, A. and LEE, S.-M. (1993). Estimating population size for continuous-time capture-recapture models via sample coverage. *Biom. J.* **35** 29–45. MR1212195 https://doi.org/10.1002/bimj.4710350104

CHOQUET, R. (2018). Markov-modulated Poisson processes as a new framework for analysing capture-recapture data. *Methods Ecol. Evol.* **9** 929–935.

CHOQUET, R., GARNIER, A., AWUVE, E. and BESNARD, A. (2017). Transient state estimation using continuous-time processes applied to opportunistic capture-recapture data. *Ecol. Model.* **361** 157–163.

CONN, P. B. and COOCH, E. G. (2009). Multistate capture-recapture analysis under imperfect state observation: An application to disease models. *J. Appl. Ecol.* **46** 486–492.

COX, D. R. and MILLER, H. D. (1965). *The Theory of Stochastic Processes*. Wiley, New York. MR0192521

DISTILLER, G. and BORCHERS, D. L. (2015). A spatially explicit capture-recapture estimator for single-catch traps. *Ecol. Evol.* **5** 5075–5087. https://doi.org/10.1002/ece3.1748

DUPUIS, J. A. (1995). Bayesian estimation of movement and survival probabilities from capture-recapture data. *Biometrika* **82** 761–772. MR1380813 https://doi.org/10.1093/biomet/82.4.761

FADDY, M. J. (1976). A note on the general time-dependent stochastic compartmental model. *Biometrics* **32** 443–448. MR0429163 https://doi.org/10.2307/2529513

FAUSTINO, C. R., JENNELLE, C. S., CONNOLLY, V., DAVIS, A. K., SWARTHOUT, E. C., DHONDT, A. A. and COOCH, E. G. (2004). Mycoplasma gallisepticum infection dynamics in a house finch population: Seasonal variation in survival, encounter and transmission rate. *J. Anim. Ecol.* **73** 651–669.

FOUCHET, D., SANTIN-JANIN, H., SAUVAGE, F., YOCCOZ, N. G. and PONTIER, D. (2016). An R package for analysing survival using continuous-time open capture-recapture models. *Methods Ecol. Evol.* **7** 518–528.

GIMENEZ, O., ROSSI, V., CHOQUET, R., DEHAIS, C., DORIS, B., VARELLA, H., VILA, J.-P. and PRADEL, R. (2007). State-space modelling of data on marked individuals. *Ecol. Model.* **206** 431–438.

GIMENEZ, O., LEBRETON, J.-D., GAILLARD, J.-M., CHOQUET, R. and PRADEL, R. (2012). Estimating demographic parameters using hidden process dynamic models. *Theor. Popul. Biol.* **82** 307–316.

GLENNIE, R., BUCKLAND, S. T., LANGROCK, R., GERRODETTE, T., BALLANCE, L. T., CHIVERS, S. J. and SCOTT, M. D. (2021). Incorporating animal movement into distance sampling. *J. Amer. Statist. Assoc.* **116** 107–115. MR4227678 https://doi.org/10.1080/01621459.2020.1764362

HWANG, W.-H. and CHAO, A. (2002). Continuous-time capture-recapture models with covariates. *Statist. Sinica* **12** 1115–1131. MR1947066

JACKSON, C. H., SHARPLES, L. D., THOMPSON, S. G., DUFFY, S. W. and COUTO, E. (2003). Multistate Markov models for disease progression with classification error. *Statistician* **52** 193–209. MR1977260 https://doi.org/10.1111/1467-9884.00351

JONSEN, I. D., FLEMMING, J. M. and MYERS, R. A. (2005). Robust state-space modeling of animal movement data. *Ecology* **86** 2874–2880.

JULLIARD, R., STENSETH, N. C., GJØSÆTER, J., LEKVE, K., FROMENTIN, J.-M. and DANIELSSEN, D. S. (2001). Natural mortality and fishing mortality in a coastal cod population: A release-recapture experiment. *Ecol. Appl.* **11** 540–558.

KAY, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics* **42** 855–865.

KENDALL, W. L. (2004). Coping with unobservable and mis–classified states in capture–recapture studies. *Anim. Biodivers. Conserv.* **27** 97–107.

KENDALL, W. L. and NICHOLS, J. D. (2002). Estimating state-transition probabilities for unobservable states using capture-recapture/resighting data. *Ecology* **83** 3276–3284.

KENDALL, W. L., NICHOLS, J. D. and HINES, J. E. (1997). Estimating temporary emigration using capture–recapture data with Pollock's robust design. *Ecology* **78** 563–578.

KING, R. (2014). Statistical ecology. *Annu. Rev. Stat. Appl.* **1** 401–426.

KING, R. and BROOKS, S. P. (2002). Bayesian model discrimination for multiple strata capture-recapture data. *Biometrika* **89** 785–806. MR1946510 https://doi.org/10.1093/biomet/89.4.785

KING, R. and BROOKS, S. P. (2003a). Closed-form likelihoods for Arnason–Schwarz models. *Biometrika* **90** 435–444. MR1986658 https://doi.org/10.1093/biomet/90.2.435

KING, R. and BROOKS, S. P. (2003b). Survival and spatial fidelity of mouflon: The effect of location, age and sex. *J. Agric. Biol. Environ. Stat.* **8** 486–513.

KING, R. and LANGROCK, R. (2016). Semi-Markov Arnason–Schwarz models. *Biometrics* **72** 619–628. MR3515788 https://doi.org/10.1111/biom.12446

KING, R. and MCCREA, R. S. (2014). A generalised likelihood framework for partially observed capture-recapture-recovery models. *Stat. Methodol.* **17** 30–45. MR3133584 https://doi.org/10.1016/j.stamet.2013.07.004

KING, R., MORGAN, B. J. T., GIMENEZ, O. and BROOKS, S. P. (2009). *Bayesian Analysis for Population Ecology*. CRC Press/CRC, Boca Raton, FL.

KOLEV, A., ROSS, G., BORCHERS, D. L. and KING, R. (2021). Spatially explicit capture-recapture as a self-exciting point process. Technical Report, Univ. College of London.

KRISHNAMURTHY, V., LEOFF, E. and SASS, J. (2018). Filterbased stochastic volatility in continuous-time hidden Markov models. *Econom. Stat.* **6** 1–21. MR3797971 https://doi.org/10.1016/j.ecosta.2016.10.007

LANGROCK, R., BORCHERS, D. L. and SKAUG, H. J. (2013). Markov-modulated nonhomogeneous Poisson processes for modeling detections in surveys of marine mammal abundance. *J. Amer. Statist. Assoc.* **108** 840–851. MR3174667 https://doi.org/10.1080/01621459.2013.797356

LANGROCK, R. and KING, R. (2013). Maximum likelihood estimation of mark-recapture-recovery models in the presence of continuous covariates. *Ann. Appl. Stat.* **7** 1709–1732. MR3127965 https://doi.org/10.1214/13-AOAS644

LEBOVIC, G. (2011). Estimating Non-Homogeneous Intensity Matrices in Continuous Time Multi-State Markov Models. PhD thesis, Univ. Toronto.

MCCLINTOCK, B. T., JOHNSON, D. S., HOOTEN, M. B., VER HOEF, J. M. and MORALES, J. M. (2014). When to be discrete: The importance of time formulation in understanding animal movement. *Mov. Ecol.* **2** 21. https://doi.org/10.1186/s40462-014-0021-6

MCCREA, R. S., MORGAN, B. J. T., GIMENEZ, O., BESBEAS, P., LEBRETON, J.-D. and BREGNBALLE, T. (2010). Multi-site integrated population modelling. *J. Agric. Biol. Environ. Stat.* **15** 539–561. MR2788639 https://doi.org/10.1007/s13253-010-0027-5Mews et al.

MEWS, S., LANGROCK, R., KING, R. and QUICK, N. (2022). Supplement to "Multistate capture–recapture models for irregularly sampled data." https://doi.org/10.1214/21-AOAS1528SUPPA, https://doi.org/10.1214/21-AOAS1528SUPPB

MICHELOT, T. and BLACKWELL, P. G. (2019). State-switching continuous-time correlated random walks. *Methods Ecol. Evol.* **10** 637–649.

NICHOLS, J. D., KENDALL, W. L., HINES, J. S. and SPENDELOW, J. A. (2004). Estimation of sex-specific survival from capture-recapture data when sex is not always known. *Ecology* **85** 3192–3201.

ORAVECZ, Z., TUERLINCKX, F. and VANDEKERCKHOVE, J. (2011). A hierarchical latent stochastic differential equation model for affective dynamics. *Psychol. Methods* **16** 468–490.

PATTERSON, T. A., PARTON, A., LANGROCK, R., BLACKWELL, P. G., THOMAS, L. and KING, R. (2017). Statistical modelling of individual animal movement: An overview of key methods and a discussion of practical challenges. *AStA Adv. Stat. Anal.* **101** 399–438. MR3712406 https://doi.org/10.1007/s10182-017-0302-7

PRADEL, R. (2005). Multievent: An extension of multistate capture-recapture models to uncertain states. *Biometrics* **61** 442–447. MR2140915 https://doi.org/10.1111/j.1541-0420.2005.00318.x

PRADEL, R. and LEBRETON, J.-D. (1999). Comparison of different approaches to study the local recruitment of breeders. *Bird Study* **46** 74–81.

ROYLE, J. A. (2008). Modeling individual effects in the Cormack–Jolly–Seber model: A state-space formulation. *Biometrics* **64** 364–370, 664. MR2432405 https://doi.org/10.1111/j.1541-0420.2007.00891.x

RUSSELL, J. R. and ENGLE, R. F. (2005). A discrete-state continuous-time model of financial transactions prices and times: The autoregressive conditional multinomial—autoregressive conditional duration model. *J. Bus. Econom. Statist.* **23** 166–180. MR2157268 https://doi.org/10.1198/073500104000000541

SCHAUB, M., GIMENEZ, O., SCHMIDT, B. R. and PRADEL, R. (2004). Estimating survival and temporary emigration in the multistate capture-recapture framework. *Ecology* **85** 2107–2113.

SCHOFIELD, M. R. and BARKER, R. J. (2008). A unified capture-recapture framework. *J. Agric. Biol. Environ. Stat.* **13** 458–477. MR2590940 https://doi.org/10.1198/108571108X383465

SCHOFIELD, M. R., BARKER, R. J. and GELLING, N. (2018). Continuous-time capture-recapture in closed populations. *Biometrics* **74** 626–635. MR3825349 https://doi.org/10.1111/biom.12763

SCHWARZ, C. J., SCHWEIGERT, J. F. and ARNASON, A. N. (1993). Estimating migration rates using tag-recovery data. *Biometrics* **49** 177–193.

VAN BEEST, F. M., MEWS, S., ELKENKAMP, S., SCHUHMANN, P., TSOLAK, D., WOBBE, T., BARTOLINO, V., BASTARDIE, F., DIETZ, R. et al. (2019). Classifying grey seal behaviour in relation to environmental variability and commercial fishing activity—a multivariate hidden Markov model. *Sci. Rep.* **9** 5642. https://doi.org/10.1038/s41598-019-42109-w

VOELKLE, M. C., OUD, J. H. L., DAVIDOV, E. and SCHMIDT, P. (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychol. Methods* **17** 176–192.

WILLIAMS, J. P., STORLIE, C. B., THERNEAU, T. M., JACK, C. R. JR. and HANNIG, J. (2020). A Bayesian approach to multistate hidden Markov models: Application to dementia progression. *J. Amer. Statist. Assoc.* **115** 16–31. MR4078442 https://doi.org/10.1080/01621459.2019.1594831

WORTHINGTON, H., MCCREA, R., KING, R. and GRIFFITHS, R. (2019). Estimating abundance from multiple sampling capture-recapture data via a multi-state multi-period stopover model. *Ann. Appl. Stat.* **13** 2043–2064. MR4037421 https://doi.org/10.1214/19-aoas1264

WÜRSIG, B. and WÜRSIG, M. (1977). The photographic determination of group size, composition and stability of coastal porpoises (Tursiops truncatus). *Science* **198** 755–756.

YIP, P. S. F., HUGGINS, R. M. and LIN, D. Y. (1996). Inference for capture-recapture experiments in continuous time with variable capture rates. *Biometrika* **83** 477–483.

YIP, P. S. F. and WANG, Y. (2002). A unified parametric regression model for recapture studies with random removals in continuous time. *Biometrics* **58** 192–199. MR1891379 https://doi.org/10.1111/j.0006-341X.2002.00192.x

ZUCCHINI, W., MACDONALD, I. L. and LANGROCK, R. (2016). *Hidden Markov Models for Time Series*: *An Introduction Using R*, 2nd ed. CRC Press, Boca Raton, FL. MR3618333

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 999–1013
https://doi.org/10.1214/21-AOAS1529

BAYESIAN INVERSE REINFORCEMENT LEARNING FOR COLLECTIVE ANIMAL MOVEMENT

BY TORYN L. J. SCHAFER[1,a], CHRISTOPHER K. WIKLE[1,b] AND MEVIN B. HOOTEN[2,3,c]

[1]*Department of Statistics, University of Missouri,* [a]*tls255@cornell.edu,* [b]*wiklec@missouri.edu*
[2]*U. S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit*
[3]*Departments of Fish, Wildlife, and Conservation Biology and Statistics, Colorado State University,* [c]*hooten@rams.colostate.edu*

Agent-based methods allow for defining simple rules that generate complex group behaviors. The governing rules of such models are typically set a priori, and parameters are tuned from observed behavior trajectories. Instead of making simplifying assumptions across all anticipated scenarios, inverse reinforcement learning provides inference on the short-term (local) rules governing long-term behavior policies by using properties of a Markov decision process. We use the computationally efficient linearly-solvable Markov decision process to learn the local rules governing collective movement for a simulation of the selfpropelled-particle (SPP) model and a data application for a captive guppy population. The estimation of the behavioral decision costs is done in a Bayesian framework with basis function smoothing. We recover the true costs in the SPP simulation and find the guppies value collective movement more than targeted movement toward shelter.

REFERENCES

ARORA, S. and DOSHI, P. (2021). A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence* **297** 103500. MR4241224 https://doi.org/10.1016/j.artint.2021.103500

BELLMAN, R. (1957). *Dynamic Programming*. Princeton Univ. Press, Princeton, NJ. MR0090477

BODE, N. W., FRANKS, D. W., WOOD, A. J., PIERCY, J. J., CROFT, D. P. and CODLING, E. A. (2012). Distinguishing social from nonsocial navigation in moving animal groups. *Amer. Nat.* **179** 621–632.

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

CHOI, J. and KIM, K.-E. (2011). Map inference for Bayesian inverse reinforcement learning. In *Advances in Neural Information Processing Systems* 1989–1997.

CHOI, J. and KIM, K.-E. (2014). Hierarchical Bayesian inverse reinforcement learning. *IEEE Trans. Cybern.* **45** 793–805.

COUZIN, I. D., KRAUSE, J., JAMES, R., RUXTON, G. D. and FRANKS, N. R. (2002). Collective memory and spatial sorting in animal groups. *J. Theoret. Biol.* **218** 1–11. MR2027139 https://doi.org/10.1006/jtbi.2002.3065

DVIJOTHAM, K. and TODOROV, E. (2010). Inverse optimal control with linearly-solvable MDPs. In *ICML* 335–342.

EARLE, A. C., SAXE, A. M. and ROSMAN, B. (2018). Hierarchical subtask discovery with non-negative matrix factorization. In *International Conference on Learning Representations*.

FINN, C., LEVINE, S. and ABBEEL, P. (2016). Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning* 49–58.

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

HOFFMAN, M. D. and GELMAN, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15** 1593–1623. MR3214779

HOOTEN, M. B., SCHARF, H. R. and MORALES, J. M. (2019). Running on empty: Recharge dynamics from animal movement data. *Ecol. Lett.* **22** 377–389. https://doi.org/10.1111/ele.13198

HOOTEN, M., WIKLE, C. and SCHWOB, M. (2020). Statistical implementations of agent-based demographic models. *Int. Stat. Rev.* **88** 441–461. MR4176185 https://doi.org/10.1111/insr.12399

Key words and phrases. Agent-based model, inverse optimal control, Markov decision process, variational approximation.

HOOTEN, M. B., JOHNSON, D. S., HANKS, E. M. and LOWRY, J. H. (2010). Agent-based inference for animal movement and selection. *J. Agric. Biol. Environ. Stat.* **15** 523–538. MR2788638 https://doi.org/10.1007/s13253-010-0038-2

HOOTEN, M. B., JOHNSON, D. S., MCCLINTOCK, B. T. and MORALES, J. M. (2017). *Animal Movement: Statistical Models for Telemetry Data*. Chapman and Hall/CRC, Boca Raton, FL.

HOOTEN, M. B., LU, X., GARLICK, M. J. and POWELL, J. A. (2020). Animal movement models with mechanistic selection functions. *Spat. Stat.* **37** 100406. MR4109595 https://doi.org/10.1016/j.spasta.2019.100406

JIN, M., DAMIANOU, A., ABBEEL, P. and SPANOS, C. (2017). Inverse reinforcement learning via deep Gaussian process. In *Conference on Uncertainty in Artificial Intelligence*.

KANGASRÄÄSIÖ, A. and KASKI, S. (2018). Inverse reinforcement learning from summary data. *Mach. Learn.* **107** 1517–1535. MR3835277 https://doi.org/10.1007/s10994-018-5730-4

KOHJIMA, M., MATSUBAYASHI, T. and SAWADA, H. (2017). Generalized inverse reinforcement learning with linearly solvable MDP. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 373–388. Springer, Berlin.

KUCUKELBIR, A., RANGANATH, R., GELMAN, A. and BLEI, D. (2015). Automatic variational inference in Stan. In *Advances in Neural Information Processing Systems* 568–576.

LEE, K., RUCKER, M., SCHERER, W. T., BELING, P. A., GERBER, M. S. and KANG, H. (2017). Agent-based model construction using inverse reinforcement learning. In 2017 *Winter Simulation Conference* (*WSC*) 1264–1275. IEEE.

MCDERMOTT, P. L., WIKLE, C. K. and MILLSPAUGH, J. (2017). Hierarchical nonlinear spatio-temporal agent-based models for collective animal movement. *J. Agric. Biol. Environ. Stat.* **22** 294–312. MR3692466 https://doi.org/10.1007/s13253-017-0289-2

MILLS FLEMMING, J. E., FIELD, C. A., JAMES, M. C., JONSEN, I. D. and MYERS, R. A. (2006). How well can animals navigate? Estimating the circle of confusion from tracking data. *Environmetrics* **17** 351–362. MR2239677 https://doi.org/10.1002/env.774

NG, A. Y. and RUSSELL, S. J. (2000). Algorithms for inverse reinforcement learning. In *ICML* 663–670.

PINSLER, R., MAAG, M., ARENZ, O. and NEUMANN, G. (2018). Inverse reinforcement learning of bird flocking behavior. *ICRA Swarms Workshop*.

RAMACHANDRAN, D. and AMIR, E. (2007). Bayesian inverse reinforcement learning. In *IJCAI* **7** 2586–2591.

RATLIFF, N. D., BAGNELL, J. A. and ZINKEVICH, M. A. (2006). Maximum margin planning. In *Proceedings of the* 23*rd International Conference on Machine Learning* 729–736.

RIED, K., MÜLLER, T. and BRIEGEL, H. J. (2019). Modelling collective motion based on the principle of agency: General framework and the case of marching locusts. *PLoS ONE* **14** e0212044. https://doi.org/10.1371/journal.pone.0212044

RUSSELL, J. C., HANKS, E. M. and HARAN, M. (2016). Dynamic models of animal movement with spatial point process interactions. *J. Agric. Biol. Environ. Stat.* **21** 22–40. MR3459292 https://doi.org/10.1007/s13253-015-0219-0

SCHAFER, T. L., WIKLE, C. K. and HOOTEN, M. B. (2022). Supplement to "Bayesian inverse reinforcement learning for collective animal movement." https://doi.org/10.1214/21-AOAS1529SUPPA, https://doi.org/10.1214/21-AOAS1529SUPPB

SCHARF, H. R., HOOTEN, M. B., FOSDICK, B. K., JOHNSON, D. S., LONDON, J. M. and DURBAN, J. W. (2016). Dynamic social networks based on movement. *Ann. Appl. Stat.* **10** 2182–2202. MR3592053 https://doi.org/10.1214/16-AOAS970

SCHARF, H. R., HOOTEN, M. B., JOHNSON, D. S. and DURBAN, J. W. (2018). Process convolution approaches for modeling interacting trajectories. *Environmetrics* **29** e2487. MR3799912 https://doi.org/10.1002/env.2487

SOSIC, A., ZOUBIR, A. M. and KOEPPL, H. (2018). A Bayesian approach to policy recognition and state representation learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **40** 1295–1308. https://doi.org/10.1109/TPAMI.2017.2711024

SOŠIĆ, A., KHUDABUKHSH, W. R., ZOUBIR, A. M. and KOEPPL, H. (2017). Inverse reinforcement learning in swarm systems. In *Proceedings of the* 16*th Conference on Autonomous Agents and MultiAgent Systems* 1413–1421.

STAN DEVELOPMENT TEAM (2020). RStan: The R interface to Stan. R package version 2.19.3.

SUTTON, R. S. and BARTO, A. G. (1998). *Introduction to Reinforcement Learning* **2**. MIT Press, Cambridge.

TODOROV, E. (2007). Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems* 1369–1376.

TODOROV, E. (2009). Efficient computation of optimal actions. *Proc. Natl. Acad. Sci. USA* **106** 11478–11483.

VICSEK, T., CZIRÓK, A., BEN-JACOB, E., COHEN, I. and SHOCHET, O. (1995). Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Lett.* **75** 1226–1229. MR3363421 https://doi.org/10.1103/PhysRevLett.75.1226

WIKLE, C. K. and HOOTEN, M. B. (2016). Hierarchical agent-based spatio-temporal dynamic models for discrete-valued data. In *Handbook of Discrete-Valued Time Series*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 349–365. CRC Press, Boca Raton, FL. MR3699413

WULFMEIER, M., ONDRUSKA, P. and POSNER, I. (2015). Deep inverse reinforcement learning. ArXiv Preprint. Available at arXiv:1507.04888.

YAMAGUCHI, S., NAOKI, H., IKEDA, M., TSUKADA, Y., NAKANO, S., MORI, I. and ISHII, S. (2018). Identification of animal behavioral strategies by inverse reinforcement learning. *PLoS Comput. Biol.* **14** e1006122. https://doi.org/10.1371/journal.pcbi.1006122

ZAMMIT-MANGION, A. (2020). FRK: Fixed Rank Kriging. R package version 0.2.2.1.

ZIEBART, B. D., MAAS, A., BAGNELL, J. A. and DEY, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence*. *AAAI'08* **3** 1433–1438. AAAI Press, Menlo Park.

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1014–1037
https://doi.org/10.1214/21-AOAS1530

A FLEXIBLE SENSITIVITY ANALYSIS APPROACH FOR UNMEASURED CONFOUNDING WITH MULTIPLE TREATMENTS AND A BINARY OUTCOME WITH APPLICATION TO SEER-MEDICARE LUNG CANCER DATA

BY LIANGYUAN HU[1,a], JUNGANG ZOU[2,b], CHENYANG GU[3,c], JIAYI JI[4,d], MICHAEL LOPEZ[5,e] AND MINAL KALE[6,f]

[1]*Department of Biostatistics and Epidemiology, Rutgers University,* [a]*liangyuan.hu@rutgers.edu*
[2]*Department of Biostatistics, Columbia University,* [b]*jz3183@cumc.columbia.edu*
[3]*Department of Biostatistics, Brown University,* [c]*chenyang.gu@alumni.brown.edu*
[4]*Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai,* [d]*jiayi.ji@mountsinai.org*
[5]*Department of Mathematics, Skidmore College,* [e]*mlopez1@skidmore.edu*
[6]*Department of Medicine, Icahn School of Medicine at Mount Sinai,* [f]*minal.kale@mountsinai.org*

In the absence of a randomized experiment, a key assumption for drawing causal inference about treatment effects is the ignorable treatment assignment. Violations of the ignorability assumption may lead to biased treatment effect estimates. Sensitivity analysis helps gauge how causal conclusions will be altered in response to the potential magnitude of departure from the ignorability assumption. However, sensitivity analysis approaches for unmeasured confounding in the context of multiple treatments and binary outcomes are scarce. We propose a flexible Monte Carlo sensitivity analysis approach for causal inference in such settings. We first derive the general form of the bias introduced by unmeasured confounding, with emphasis on theoretical properties uniquely relevant to multiple treatments. We then propose methods to encode the impact of unmeasured confounding on potential outcomes and adjust the estimates of causal effects in which the presumed unmeasured confounding is removed. Our proposed methods embed nested multiple imputation within the Bayesian framework, which allow for seamless integration of the uncertainty about the values of the sensitivity parameters and the sampling variability as well as use of the Bayesian Additive Regression Trees for modeling flexibility. Expansive simulations validate our methods and gain insight into sensitivity analysis with multiple treatments. We use the SEER-Medicare data to demonstrate sensitivity analysis using three treatments for early stage nonsmall cell lung cancer. The methods developed in this work are readily available in the R package SAMTx.

REFERENCES

BILLÉ, A., BUXTON, J., VIVIANO, A., GAMMON, D., VERES, L., ROUTLEDGE, T., HARRISON-PHIPPS, K., DIXON, A. and MINETTO, M. A. (2021). Preoperative physical activity predicts surgical outcomes following lung cancer resection. *Integr. Cancer Ther.* **20** 1–8.

BRUMBACK, B. A., HERNÁN, M. A., HANEUSE, S. J. and ROBINS, J. M. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Stat. Med.* **23** 749–767.

CEPPA, D. P., KOSINSKI, A. S., BERRY, M. F., TONG, B. C., HARPOLE, D. H., MITCHELL, J. D., D'AMICO, T. A. and ONAITIS, M. W. (2012). Thoracoscopic lobectomy has increasing benefit in patients with poor pulmonary function: A Society of Thoracic Surgeons Database analysis. *Ann. Surg.* **256** 487–493. https://doi.org/10.1097/SLA.0b013e318265819c

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

Key words and phrases. Causal inference, ignorability assumption, observational data, Bayesian inference, nested multiple imputation.

DANIELS, M. J. and HOGAN, J. W. (2008). *Missing Data in Longitudinal Studies*: *Strategies for Bayesian Modeling and Sensitivity Analysis*. *Monographs on Statistics and Applied Probability* **109**. CRC Press/CRC, Boca Raton, FL. MR2459796 https://doi.org/10.1201/9781420011180

DING, P. and VANDERWEELE, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27** 368–377. https://doi.org/10.1097/EDE.0000000000000457

DORIE, V., HARADA, M., CARNEGIE, N. B. and HILL, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Stat. Med.* **35** 3453–3470. MR3537215 https://doi.org/10.1002/sim.6973

GREENLAND, S. (2005). Multiple-bias modelling for analysis of observational data. *J. Roy. Statist. Soc. Ser. A* **168** 267–306. MR2119402 https://doi.org/10.1111/j.1467-985X.2004.00349.x

GU, C. and GUTMAN, R. (2019). Development of a common patient assessment scale across the continuum of care: A nested multiple imputation approach. *Ann. Appl. Stat.* **13** 466–491. MR3937437 https://doi.org/10.1214/18-AOAS1202

GUSTAFSON, P. and MCCANDLESS, L. C. (2018). When is a sensitivity parameter exactly that? *Statist. Sci.* **33** 86–95. MR3757506 https://doi.org/10.1214/17-STS632

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

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

HOGAN, J. W., DANIELS, M. J. and HU, L. (2014). A Bayesian perspective on assessing sensitivity to assumptions about unobserved data. In *Handbook of Missing Data Methodology* (G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis and G. Verbeke, eds.) 18, 405–434. CRC Press, Boca Raton, FL.

HOWINGTON, J. A., BLUM, M. G., CHANG, A. C., BALEKIAN, A. A. and MURTHY, S. C. (2013). Treatment of stage I and II non-small cell lung cancer: Diagnosis and management of lung cancer: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* **143** e278S–e313S.

HU, L. (2020). Discussion on "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects" by Hahn, Murray and Carvalho. *Bayesian Anal.* **15** 1020–1023.

HU, L. and GU, C. (2020). Estimation of causal effects of multiple treatments in healthcare database studies with rare outcomes. *Health Serv. Outcomes Res. Methodol.* **21** 287–308.

HU, L. and HOGAN, J. W. (2019). Causal comparative effectiveness analysis of dynamic continuous-time treatment initiation rules with sparsely measured outcomes and death. *Biometrics* **75** 695–707. MR3999191

HU, L., JI, J. and LI, F. (2021). Estimating heterogeneous survival treatment effect in observational data using machine learning. *Stat. Med.* **40** 4691–4713.

HU, L., LIN, J.-Y. J. and JI, J. (2021). Variable selection with missing data in both covariates and outcomes: Imputation and machine learning. *Stat. Methods Med. Res.* To appear.

HU, L., LIU, B. and LI, Y. (2020). Ranking sociodemographic, health behavior, prevention, and environmental factors in predicting neighborhood cardiovascular health: A Bayesian machine learning approach. *Prev. Med.* **141** 106240.

HU, L., HOGAN, J. W., MWANGI, A. W. and SIIKA, A. (2018). Modeling the causal effect of treatment initiation time on survival: Application to HIV/TB co-infection. *Biometrics* **74** 703–713. MR3825357 https://doi.org/10.1111/biom.12780

HU, L., GU, C., LOPEZ, M., JI, J. and WISNIVESKY, J. (2020a). Estimation of causal effects of multiple treatments in observational studies with a binary outcome. *Stat. Methods Med. Res.* **29** 3218–3234. MR4156850 https://doi.org/10.1177/0962280220921909

HU, L., LIU, B., JI, J. and LI, Y. (2020b). Tree-based machine learning to identify and understand major determinants for stroke at the neighborhood level. *J. Am. Heart Assoc.* **9** e016745.

HU, L., LIN, J.-Y. J., SIGEL, K. and KALE, M. (2021a). Estimating heterogeneous survival treatment effects of lung cancer screening approaches: A causal machine learning analysis. *Ann. Epidemiol.* **62** 36–42.

HU, L., ZOU, J., GU, C., JI, J., LOPEZ, M. and KALE, M. (2022). Supplement to "A flexible sensitivity analysis approach for unmeasured confounding with multiple treatments and a binary outcome with application to SEER-Medicare lung cancer data." https://doi.org/10.1214/21-AOAS1530SUPPA, https://doi.org/10.1214/21-AOAS1530SUPPB

IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* **93** 126–132.

KASZA, J., WOLFE, R. and SCHUSTER, T. (2017). Assessing the impact of unmeasured confounding for binary outcomes using confounding functions. *Int. J. Epidemiol.* **46** 1303–1311. https://doi.org/10.1093/ije/dyx023

LAKENS, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Front. Psychol.* **4** 863. https://doi.org/10.3389/fpsyg.2013.00863

LASH, T. L., FOX, M. P. and FINK, A. K. (2011). *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer Science & Business Media, New York.

LI, L., SHEN, C., WU, A. C. and LI, X. (2011). Propensity score-based sensitivity analysis method for uncontrolled confounding. *Am. J. Epidemiol.* **174** 345–353.

LIN, D. Y., PSATY, B. M. and KRONMAL, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54** 948–963.

MCCANDLESS, L. C. and GUSTAFSON, P. (2017). A comparison of Bayesian and Monte Carlo sensitivity analysis for unmeasured confounding. *Stat. Med.* **36** 2887–2901. MR3670397 https://doi.org/10.1002/sim.7298

ROBINS, J. M. (1999). Association, causation, and marginal structural models. *Synthese* **121** 151–179. MR1766776 https://doi.org/10.1023/A:1005285815569

ROSENBAUM, P. R. and RUBIN, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Roy. Statist. Soc. Ser. B* **45** 212–218.

RUAN, A. and KULKARNI, V. (2020). Anesthesia considerations for robotic thoracic surgery. *Video-Assist. Thorac. Surg.* **5** 1–8.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

RUBIN, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Stat. Neerl.* **57** 3–18. MR2055518 https://doi.org/10.1111/1467-9574.00217

SAITO, H., HATAKEYAMA, K., KONNO, H., MATSUNAGA, T., SHIMADA, Y. and MINAMIYA, Y. (2017). Impact of pulmonary rehabilitation on postoperative complications in patients with lung cancer and chronic obstructive pulmonary disease. *Thorac. Cancer* **8** 451–460. https://doi.org/10.1111/1759-7714.12466

SIHOE, A. D. L. (2020). Video-assisted thoracoscopic surgery as the gold standard for lung cancer surgery. *Respirology* **25** 49–60. https://doi.org/10.1111/resp.13920

VANDERWEELE, T. J. and ARAH, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* **22** 42–52. https://doi.org/10.1097/EDE.0b013e3181f74493

VON ELM, E., ALTMAN, D. G., EGGER, M., POCOCK, S. J., GØTZSCHE, P. C. and VANDENBROUCKE, J. P. (2007). The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Ann. Intern. Med.* **147** 573–577.

ZHOU, X. and REITER, J. P. (2010). A note on Bayesian inference after multiple imputation. *Amer. Statist.* **64** 159–163. MR2757007 https://doi.org/10.1198/tast.2010.09109

ZIGLER, C. M., WATTS, K., YEH, R. W., WANG, Y., COULL, B. A. and DOMINICI, F. (2013). Model feedback in Bayesian propensity score estimation. *Biometrics* **69** 263–273. MR3058073 https://doi.org/10.1111/j.1541-0420.2012.01830.x

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1038–1070
https://doi.org/10.1214/21-AOAS1531

ROBUST BAYESIAN INFERENCE FOR BIG DATA: COMBINING SENSOR-BASED RECORDS WITH TRADITIONAL SURVEY DATA

BY ALI RAFEI[1,a], CAROL A. C. FLANNAGAN[2,b], BRADY T. WEST[3,c] AND MICHAEL R. ELLIOTT[4,d]

[1]*Program in Survey and Data Science, University of Michigan,* [a]*arafei@umich.edu*
[2]*University of Michigan Transportation Research Institute,* [b]*cacf@umich.edu*
[3]*Program in Survey and Data Science, University of Michigan,* [c]*bwest@umich.edu*
[4]*Program in Survey and Data Science, Department of Biostatistics, University of Michigan,* [d]*mrelliot@umich.edu*

Big Data often presents as massive nonprobability samples. Not only is the selection mechanism often unknown but larger data volume amplifies the relative contribution of selection bias to total error. Existing bias adjustment approaches assume that the conditional mean structures have been correctly specified for the selection indicator or key substantive measures. In the presence of a reference probability sample, these methods rely on a pseudolikelihood method to account for the sampling weights of the reference sample, which is parametric in nature. Under a Bayesian framework, handling the sampling weights is an even bigger hurdle. To further protect against model misspecification, we expand the idea of double robustness such that more flexible nonparametric methods as well as Bayesian models can be used for prediction. In particular, we employ Bayesian additive regression trees which not only capture nonlinear associations automatically but permit direct quantification of the uncertainty of point estimates through its posterior predictive draws. We apply our method to sensor-based naturalistic driving data from the second Strategic Highway Research Program using the 2017 National Household Travel Survey as a benchmark.

REFERENCES

AN, W. (2010). Bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. *Sociol. Method.* **40** 151–189.

AN, H. and LITTLE, R. J. A. (2008). Robust model-based inference for incomplete data via penalized spline propensity prediction. *Comm. Statist. Simulation Comput.* **37** 1718–1731. MR2542428 https://doi.org/10.1080/03610910802255840

ANTIN, J., STULCE, K., EICHELBERGER, L. and HANKEY, J. (2015). Naturalistic driving study: Descriptive comparison of the study sample with national data. Technical Report.

BAKER, R., BRICK, J. M., BATES, N. A., BATTAGLIA, M., COUPER, M. P., DEVER, J., GILE, K. J. and TOURANGEAU, R. (2013). Summary report of the AAPOR Task Force on non-probability sampling. *Journal of Survey Statistics and Methodology* **1** 90–143.

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

CAMPBELL, K. L. (2012). The SHRP 2 naturalistic driving study: Addressing driver performance and behavior in traffic safety. *TR News* **282** 30–35.

CASTANEDO, F. (2013). A review of data fusion techniques. *The Scientific World Journal*.

CHEN, Y., LI, P. and WU, C. (2020). Doubly robust inference with nonprobability survey samples. *J. Amer. Statist. Assoc.* **115** 2011–2021. MR4189773 https://doi.org/10.1080/01621459.2019.1677241

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2007). Bayesian ensemble learning. In *Advances in Neural Information Processing Systems* 265–272.

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

Key words and phrases. Big Data, nonprobability sample, quasi-randomization, prediction model, doubly robust, augmented inverse propensity weighting, Bayesian additive regression trees.

COUPER, M. (2013). Is the sky falling? New technology, changing media, and the future of surveys. Keynote presentation at the 5th European Survey Research Association Conference. Ljubliana, Slovenia.
DAAS, P. J., PUTS, M. J., BUELENS, B. and VAN DEN HURK, P. A. (2015). Big data as a source for official statistics. *J. Off. Stat.* **31** 249–262.
DONG, Q., ELLIOTT, M. R. and RAGHUNATHAN, T. E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Surv. Methodol.* **40** 29–46.
ELLIOTT, M. R. and VALLIANT, R. (2017). Inference for nonprobability samples. *Statist. Sci.* **32** 249–264. MR3648958 https://doi.org/10.1214/16-STS598
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
GROVES, R. M. (2011). Three eras of survey research. *Public Opin. Q.* **75** 861–871.
GUO, F., HANKEY, J. M. et al. (2009). Modeling 100-car safety events: A case-based approach for analyzing naturalistic driving data. Technical Report, Virginia Tech Transportation Institute.
HAN, P. and WANG, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika* **100** 417–430. MR3068443 https://doi.org/10.1093/biomet/ass087
HAZIZA, D. and RAO, J. N. K. (2005). Inference for domains under imputation for missing survey data. *Canad. J. Statist.* **33** 149–161. MR2193025 https://doi.org/10.1002/cjs.5550330201
HILL, J. and SU, Y.-S. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Ann. Appl. Stat.* **7** 1386–1420. MR3127952 https://doi.org/10.1214/13-AOAS630
HONG, H., RUDOLPH, K. E. and STUART, E. A. (2017). Bayesian approach for addressing differential covariate measurement error in propensity score methods. *Psychometrika* **82** 1078–1096. MR3736342 https://doi.org/10.1007/s11336-016-9533-x
HUISINGH, C., OWSLEY, C., LEVITAN, E. B., IRVIN, M. R., MACLENNAN, P. and MCGWIN, G. (2019). Distracted driving and risk of crash or near-crash involvement among older drivers using naturalistic driving data with a case-crossover study design. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* **74** 550–555.
HUNSBERGER, S., GRAUBARD, B. I. and KORN, E. L. (2008). Testing logistic regression coefficients with clustered data and few positive outcomes. *Stat. Med.* **27** 1305–1324. MR2420159 https://doi.org/10.1002/sim.3011
JAPEC, L., KREUTER, F., BERG, M., BIEMER, P., DECKER, P., LAMPE, C., LANE, J., O'NEIL, C. and USHER, A. (2015). Big data in survey research: AAPOR task force report. *Public Opin. Q.* **79** 839–880.
JOHNSON, T. P. and SMITH, T. W. (2017). Big Data and Survey Research: Supplement or Substitute?. In *Seeing Cities Through Big Data* 113–125. Springer, Berlin.
KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458 https://doi.org/10.1214/07-STS227
KAPLAN, D. and CHEN, J. (2012). A two-step Bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika* **77** 581–609. MR2943114 https://doi.org/10.1007/s11336-012-9262-8
KIM, J. K. and HAZIZA, D. (2014). Doubly robust inference with missing data in survey sampling. *Statist. Sinica* **24** 375–394. MR3183689
KIM, J. K. and PARK, H. (2006). Imputation using response probability. *Canad. J. Statist.* **34** 171–182. MR2280924 https://doi.org/10.1002/cjs.5550340112
KIM, J. K. and RAO, J. N. K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika* **99** 85–100. MR2899665 https://doi.org/10.1093/biomet/asr063
KIM, J.-K. and TAM, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *Int. Stat. Rev.*. **89** 382–401.
KIM, J. K., BRICK, J. M., FULLER, W. A. and KALTON, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 509–521. MR2278338 https://doi.org/10.1111/j.1467-9868.2006.00546.x
KIM, J. K., PARK, S., CHEN, Y. and WU, C. (2021). Combining non-probability and probability survey samples through mass imputation. *J. Roy. Statist. Soc. Ser. A* **184** 941–963. MR4305566 https://doi.org/10.1111/rssa.12696
KITCHIN, R. (2015). The opportunities, challenges and risks of big data for official statistics. *Stat. J. IAOS* **31** 471–481.
KOTT, P. S. (1994). A note on handling nonresponse in sample surveys. *J. Amer. Statist. Assoc.* **89** 693–696. MR1294093
KOTT, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Surv. Methodol.* **32** 133–142.

KOTT, P. S. and CHANG, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *J. Amer. Statist. Assoc.* **105** 1265–1275. MR2752620 https://doi.org/10.1198/jasa.2010.tm09016
KREUTER, F. and PENG, R. D. (2014). Extracting information from Big Data: Issues of measurement, inference and linkage. *Privacy, Big Data, and the Public Good: Frameworks for Engagement* 257.
LANE, J. (2016). Big data for public policy: The quadruple helix. *J. Policy Anal. Manage.* **35** 708–715.
LEE, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *J. Off. Stat.* **22** 329.
LEE, S. and VALLIANT, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociol. Methods Res.* **37** 319–343. MR2649463 https://doi.org/10.1177/0049124108329643
LITTLE, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *J. Amer. Statist. Assoc.* **99** 546–556. MR2109316 https://doi.org/10.1198/016214504000000467
LITTLE, R. J. and ZHENG, H. (2007). The Bayesian approach to the analysis of finite population surveys. *Bayesian Statistics* **8** 1–20.
MCCANDLESS, L. C., GUSTAFSON, P. and AUSTIN, P. C. (2009). Bayesian propensity score analysis for observational data. *Stat. Med.* **28** 94–112. MR2655553 https://doi.org/10.1002/sim.3460
MCGUCKIN, N. and FUCCI, A. (2018). Summary of travel trends: 2017 National household travel survey (Report FHWA-PL-18-019). Washington, DC: Federal Highway Administration, US Department of Transportation.
MENG, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* **12** 685–726. MR3834282 https://doi.org/10.1214/18-AOAS1161SF
MILLER, P. V. (2017). Is there a future for surveys? *Public Opin. Q.* **81** 205–212.
MURDOCH, T. B. and DETSKY, A. S. (2013). The inevitable application of big data to health care. *J. Am. Med. Assoc.* **309** 1351–1352.
OMAN, S. D. and ZUCKER, D. M. (2001). Modelling and generating correlated binary variables. *Biometrika* **88** 287–290. MR1841276 https://doi.org/10.1093/biomet/88.1.287
PFEFFERMANN, D. and SVERCHKOV, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya, Ser. B* **61** 166–186. MR1720710
PFEFFERMANN, D. and SVERCHKOV, M. (2009). Inference under informative sampling. In *Handbook of Statistics* **29** 455–487. Elsevier, Amsterdam.
RAFEI, A., FLANNAGAN, C. A. C. and ELLIOTT, M. R. (2020). Big Data for Finite Population Inference: Applying Quasi-random Approaches to Naturalistic Driving Data using Bayesian Additive Regression Trees. *Journal of Survey Statistics and Methodology* **8** 148–180.
RAFEI, A., FLANNAGAN, C. A, WEST, B. T and ELLIOTT, M. R (2022). Supplement to "Robust Bayesian inference for big data: Combining sensor-based records with traditional survey data." https://doi.org/10.1214/21-AOAS1531SUPPA, https://doi.org/10.1214/21-AOAS1531SUPPB
RAO, J. N. K. and WU, C.-F. J. (1988). Resampling inference with complex survey data. *J. Amer. Statist. Assoc.* **83** 231–241. MR0941020
RIVERS, D. (2007). Sampling for web surveys. In *Joint Statistical Meetings*.
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. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41
RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys. Wiley Classics Library*. Wiley Interscience, Hoboken, NJ. MR2117498
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
SANTOS, A., MCGUCKIN, N., NAKAMOTO, H. Y., GRAY, D. and LISS, S. (2011). Summary of travel trends: 2009 national household travel survey. Technical Report.
SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478 https://doi.org/10.2307/2669923
SENTHILKUMAR, S., RAI, B. K., MESHRAM, A. A., GUNASEKARAN, A. and CHANDRAKUMARMANGALAM, S. (2018). Big Data in healthcare management: A review of literature. *American Journal of Theoretical and Applied Business* **4** 57–69.
SMITH, T. M. F. (1983). On the validity of inferences from nonrandom samples. *J. Roy. Statist. Soc. Ser. A* **146** 394–403. MR0769995 https://doi.org/10.2307/2981454
STRUIJS, P., BRAAKSMA, B. and DAAS, P. J. (2014). Official statistics and big data. *Big Data & Society* **1** 1–6.
TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* **101** 1619–1637. MR2279484 https://doi.org/10.1198/016214506000000023

TAN, Y. V., ELLIOTT, M. R. and FLANNAGAN, C. A. C. (2017). Development of a real-time prediction model of driver behavior at intersections using kinematic time series data. *Accident Anal. Prev.* **106** 428–436. https://doi.org/10.1016/j.aap.2017.07.003

TAN, Y. V., FLANNAGAN, C. A. C. and ELLIOTT, M. R. (2018). Predicting human-driving behavior to help driverless vehicles drive: Random intercept Bayesian additive regression trees. *Stat. Interface* **11** 557–572. MR3858513 https://doi.org/10.4310/SII.2018.v11.n4.a1

TAN, Y. V., FLANNAGAN, C. A. and ELLIOTT, M. R. (2019). "Robust-Squared" imputation models using Bart. *Journal of Survey Statistics and Methodology* **7** 465–497.

TRANSPORTATION RESEARCH BOARD NATIONAL ACADEMY OF SCIENCES (2013). The 2nd Strategic Highway Research Program Naturalistic Driving Study Dataset.

VALLIANT, R. and DEVER, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociol. Methods Res.* **40** 105–137. MR2758301 https://doi.org/10.1177/0049124110392533

WANG, W., ROTHSCHILD, D., GOEL, S. and GELMAN, A. (2015). Forecasting elections with non-representative polls. *Int. J. Forecast.* **31** 980–991.

ZANGENEH, S. Z. and LITTLE, R. J. (2015). Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample. *Journal of Survey Statistics and Methodology* **3** 162–192.

ZHANG, G. and LITTLE, R. (2011). A comparative study of doubly robust estimators of the mean with missing data. *J. Stat. Comput. Simul.* **81** 2039–2058. MR2864188 https://doi.org/10.1080/00949655.2010.516750

ZHOU, Q., MCNEAL, C., COPELAND, L. A., ZACHARIAH, J. P. and SONG, J. J. (2020). Bayesian propensity score analysis for clustered observational data. *Stat. Methods Appl.* **29** 335–355. MR4106804 https://doi.org/10.1007/s10260-019-00484-8

ZIGLER, C. M., WATTS, K., YEH, R. W., WANG, Y., COULL, B. A. and DOMINICI, F. (2013). Model feedback in Bayesian propensity score estimation. *Biometrics* **69** 263–273. MR3058073 https://doi.org/10.1111/j.1541-0420.2012.01830.x

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1071–1089
https://doi.org/10.1214/21-AOAS1532

A SPARSE NEGATIVE BINOMIAL CLASSIFIER WITH COVARIATE ADJUSTMENT FOR RNA-SEQ DATA

BY TANBIN RAHMAN[1,a], HSIN-EN HUANG[2,d], YUJIA LI[1,b], AN-SHUN TAI[2,e], WEN-PING HSEIH[2,f], COLLEEN A. MCCLUNG[3,g] AND GEORGE TSENG[1,c]

[1]*Department of Biostatistics, University of Pittsburgh,* [a]*mdr56@pitt.edu,* [b]*yul178@pitt.edu,* [c]*ctseng@pitt.edu*
[2]*Institute of Statistics, National Tsing Hua University,* [d]*kid0857@gmail.com,* [e]*s9824509@m98.nthu.edu.tw,* [f]*wphsich@stat.nthu.edu.tw*
[3]*Deparment of Psychiatry, University of Pittsburgh,* [g]*mcclungca@upmc.edu*

Supervised machine learning methods have been increasingly used in biomedical research and clinical practice. In transcriptomic applications, RNA-seq data have become dominating and have gradually replaced traditional microarray, due to their reduced background noise and increased digital precision. Most existing machine learning methods are, however, designed for continuous intensities of microarray and are not suitable for RNA-seq count data. In this paper we develop a negative binomial model via generalized linear model framework with double regularization for gene and covariate sparsity to accommodate three key elements: adequate modeling of count data with overdispersion, gene selection and adjustment for covariate effect. The proposed sparse negative binomial classifier (snbClass) is evaluated in simulations and two real applications of multidisease postmortem brain tissue RNA-seq data and cervical tumor miRNA-seq data to demonstrate its superior performance in prediction accuracy and feature selection.

REFERENCES

ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*. **11** R106. https://doi.org/10.1186/gb-2010-11-10-r106

BALZAMINO, B. O., ESPOSITO, G., MARINO, R., KELLER, F. and MICERA, A. (2015). NGF expression in reelin-deprived retinal cells: A potential neuroprotective effect. *Neuromol. Med*. **17** 314–325.

BRADLEY, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. **30** 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

BROWN, M. P., GRUNDY, W. N., LIN, D., CRISTIANINI, N., SUGNET, C. W., FUREY, T. S., ARES, M. and HAUSSLER, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* **97** 262–267.

BULLARD, J. H., PURDOM, E., HANSEN, K. D. and DUDOIT, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinform*. **11** 94. https://doi.org/10.1186/1471-2105-11-94

CHU, Y. and COREY, D. R. (2012). RNA sequencing: Platform selection, experimental design, and data interpretation. *Nucleic. Acid Ther.* **22** 271–274. PMID: 22830413. https://doi.org/10.1089/nat.2012.0367

CONESA, A., MADRIGAL, P., TARAZONA, S., GOMEZ-CABRERO, D., CERVERA, A., MCPHERSON, A., SZCZEŚNIAK, M. W., GAFFNEY, D. J., ELO, L. L. et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol*. **17** 13.

DÍAZ-URIARTE, R. and DE ANDRES, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinform*. **7** 3.

DONG, K., ZHAO, H., TONG, T. and WAN, X. (2016). NBLDA: Negative binomial linear discriminant analysis for RNA-Seq data. *BMC Bioinform*. **17** 369. https://doi.org/10.1186/s12859-016-1208-1

DUDOIT, S., FRIDLYAND, J. and SPEED, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc*. **97** 77–87. MR1963389 https://doi.org/10.1198/016214502753479248

FROMER, M., ROUSSOS, P., SIEBERTS, S. K., JOHNSON, J. S., KAVANAGH, D. H., PERUMAL, T. M., RUDERFER, D. M., OH, E. C., TOPOL, A. et al. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci*. **19** 1442.

Key words and phrases. Classification, count data, RNA-seq data, covariate adjustment.

LI, M. D., CAO, J., WANG, S., WANG, J., SARKAR, S., VIGORITO, M., MA, J. Z. and CHANG, S. L. (2013). Transcriptome sequencing of gene expression in the brain of the HIV-1 transgenic rat. *PLoS ONE* **8** e59582.

LI, Y., RAHMAN, T., MA, T., TANG, L. and TSENG, G. C. (2021). A sparse negative binomial mixture model for clustering RNA-seq count data. *Biostatistics*.

LIM, D. K., RASHID, N. U. and IBRAHIM, J. G. (2021). Model-based feature selection and clustering of RNA-seq data for unsupervised subtype discovery. *Ann. Appl. Stat.* **15** 481–508. MR4255285 https://doi.org/10.1214/20-aoas1407

LORENZ, D. J., GILL, R. S., MITRA, R. and DATTA, S. (2014). Using RNA-seq data to detect differentially expressed genes. In *Statistical Analysis of Next Generation Sequencing Data* 25–49. Springer, Berlin.

MARIONI, J. C., MASON, C. E., MANE, S. M., STEPHENS, M. and GILAD, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18** 1509–1517.

MCCARTHY, D. J., CHEN, Y. and SMYTH, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40** 4288–4297.

PETERS, M. J., JOEHANES, R., PILLING, L. C., SCHURMANN, C., CONNEELY, K. N., POWELL, J., REINMAA, E., SUTPHIN, G. L., ZHERNAKOVA, A. et al. (2015). The transcriptional landscape of age in human peripheral blood. *Nat. Commun.* **6** 8570. https://doi.org/10.1038/ncomms9570

RAHMAN, T., HUANG, H.-E., LI, Y., TAI, A.-S., HSEIH, W.-P., MCCLUNG, C. A. and TSENG, G. (2022). Supplement to "A sparse negative binomial classifier with covariate adjustment for RNA-seq data." https://doi.org/10.1214/21-AOAS1532SUPPA, https://doi.org/10.1214/21-AOAS1532SUPPB

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

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

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36** 111–147. MR0356377

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. and CHU, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99** 6567–6572.

WANG, Z., GERSTEIN, M. and SNYDER, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10** 57–63. https://doi.org/10.1038/nrg2484

WITTEN, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *Ann. Appl. Stat.* **5** 2493–2518. MR2907124 https://doi.org/10.1214/11-AOAS493

WITTEN, D., TIBSHIRANI, R., GU, S. G., FIRE, A. and LUI, W.-O. (2010). Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biol.* **8** 58. https://doi.org/10.1186/1741-7007-8-58

ZARARSIZ, G., GOKSULUK, D., KORKMAZ, S., ELDEM, V., ZARARSIZ, G. E., DURU, I. P. and OZTURK, A. (2017). A comprehensive simulation study on classification of RNA-Seq data. *PLoS ONE* **12** e0182507.

ZHAO, S., FUNG-LEUNG, W.-P., BITTNER, A., NGO, K. and LIU, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE* **9** e78644.

ZHENG, W.-H., KAR, S., DORE, S. and QUIRION, R. (2000). Insulin-like growth factor-1 (IGF-1): A neuroprotective trophic factor acting via the Akt kinase pathway. *Adv. Res. Neurodegener.* 261–272.

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1090–1110
https://doi.org/10.1214/21-AOAS1533

KERNEL MACHINE AND DISTRIBUTED LAG MODELS FOR ASSESSING WINDOWS OF SUSCEPTIBILITY TO ENVIRONMENTAL MIXTURES IN CHILDREN'S HEALTH STUDIES

BY ANDER WILSON[1,a], HSIAO-HSIEN LEON HSU[2,b], YUEH-HSIU MATHILDA CHIU[2,c], ROBERT O. WRIGHT[2,d], ROSALIND J. WRIGHT[2,e] AND BRENT A. COULL[3,f]

[1]*Department of Statistics, Colorado State University,* [a]*ander.wilson@colostate.edu*

[2]*Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai,* [b]*leon.hsu@mssm.edu,* [c]*mathilda.chiu@mssm.edu,* [d]*robert.wright@mssm.edu,* [e]*rosalind.wright@mssm.edu*

[3]*Department of Biostatistics, Harvard T. H. Chan School of Public Health,* [f]*bcoull@hsph.harvard.edu*

Exposures to environmental chemicals during gestation can alter health status later in life. Most studies of maternal exposure to chemicals during pregnancy have focused on a single chemical exposure observed at high temporal resolution. Recent research has turned to focus on exposure to mixtures of multiple chemicals, generally observed at a single time point. We consider statistical methods for analyzing data on chemical mixtures that are observed at a high temporal resolution. As motivation, we analyze the association between exposure to four ambient air pollutants observed weekly throughout gestation and birth weight in a Boston-area prospective birth cohort. To explore patterns in the data, we first apply methods for analyzing data on: (1) a single chemical observed at high temporal resolution, and (2) a mixture measured at a single point in time. We highlight the shortcomings of these approaches for temporally-resolved data on exposure to chemical mixtures. Second, we propose a novel method, a Bayesian kernel machine regression distributed lag model (BKMR-DLM) that simultaneously accounts for nonlinear associations and interactions among time-varying measures of exposure to mixtures. BKMR-DLM uses a functional weight for each exposure that parameterizes the window of susceptibility corresponding to that exposure within a kernel machine framework that captures nonlinear and interaction effects of the multivariate exposure on the outcome. In a simulation study we show that the proposed method can better estimate the exposure-response function and, in high signal settings, can identify critical windows in time during which exposure has an increased association with the outcome. Applying the proposed method to the Boston birth cohort data, we find evidence of a negative association between organic carbon and birth weight and that nitrate modifies the organic carbon, elemental carbon, and sulfate exposure-response functions.

REFERENCES

BAUER, J. A., CLAUS HENN, B., AUSTIN, C., ZONI, S., FEDRIGHI, C., CAGNA, G., PLACIDI, D., WHITE, R. F., YANG, Q. et al. (2017). Manganese in teeth and neurobehavior: Sex-specific windows of susceptibility. *Environ. Int.* **108** 299–308. https://doi.org/10.1016/j.envint.2017.08.013

BELLO, G. A., ARORA, M., AUSTIN, C., HORTON, M. K., WRIGHT, R. O. and GENNINGS, C. (2017). Extending the distributed lag model framework to handle chemical mixtures. *Environ. Res.* **156** 253–264. https://doi.org/10.1016/j.envres.2017.03.031

BOBB, J. F. (2017). bkmr: Bayesian kernel machine regression.

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

Key words and phrases. Air pollution, chemical mixtures, children's health, windows of susceptibility, distributed lag models, kernel machine regression.

Bobb, J. F., Claus Henn, B., Valeri, L. and Coull, B. A. (2018). Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environ. Health* **17** 67. https://doi.org/10.1186/s12940-018-0413-y

Bose, S., Chiu, Y.-H. M., Hsu, H.-H. L., Di, Q., Rosa, M. J., Lee, A., Kloog, I., Wilson, A., Schwartz, J. et al. (2017). Prenatal nitrate exposure and childhood asthma. Influence of maternal prenatal stress and fetal sex. *Am. J. Respir. Crit. Care Med.* **196** 1396–1403. https://doi.org/10.1164/rccm.201702-0421OC

Braun, J. M., Gennings, C., Hauser, R. and Webster, T. F. (2016). What can epidemiological studies tell us about the impact of chemical mixtures on human health? *Environ. Health Perspect.* **124** A6–A9. https://doi.org/10.1289/ehp.1510569

Carrico, C., Gennings, C., Wheeler, D. C. and Factor-Litvak, P. (2015). Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *J. Agric. Biol. Environ. Stat.* **20** 100–120. MR3334469 https://doi.org/10.1007/s13253-014-0180-3

Chang, H. H., Reich, B. J. and Miranda, M. L. (2012). Time-to-event analysis of fine particle air pollution and preterm birth: Results from North Carolina, 2001–2005. *Am. J. Epidemiol.* **175** 91–98. https://doi.org/10.1093/aje/kwr403

Chang, H. H., Warren, J. L., Darrow, L. A., Reich, B. J. and Waller, L. A. (2015). Assessment of critical exposure and outcome windows in time-to-event analysis with application to air pollution and preterm birth study. *Biostatistics* **16** 509–521. MR3365443 https://doi.org/10.1093/biostatistics/kxu060

Chen, Y.-H., Mukherjee, B. and Berrocal, V. J. (2019). Distributed lag interaction models with two pollutants. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 79–97. MR3902983

Chen, Y.-H., Mukherjee, B., Adar, S. D., Berrocal, V. J. and Coull, B. A. (2018). Robust distributed lag models using data adaptive shrinkage. *Biostatistics* **19** 461–478. MR3867407 https://doi.org/10.1093/biostatistics/kxx041

Claus Henn, B., Austin, C., Coull, B. A., Schnaas, L., Gennings, C., Horton, M. K., Hernández-Ávila, M., Hu, H., Téllez-Rojo, M. M. et al. (2018). Uncovering neurodevelopmental windows of susceptibility to manganese exposure using dentine microspatial analyses. *Environ. Res.* **161** 588–598. https://doi.org/10.1016/j.envres.2017.12.003

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

Davalos, A. D., Luben, T. J., Herring, A. H. and Sacks, J. D. (2017). Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures. *Ann. Epidemiol.* **27** 145–153. https://doi.org/10.1016/j.annepidem.2016.11.016

Di, Q., Koutrakis, P. and Schwartz, J. (2016). A hybrid prediction model for PM2.5 mass and components using a chemical transport model and land use regression. *Atmos. Environ.* **131** 390–399. https://doi.org/10.1016/j.atmosenv.2016.02.002

Gasparrini, A. (2011). Distributed lag linear and non-linear models in R: The package dlnm. *J. Stat. Softw.* **43** 1–20. https://doi.org/10.18637/jss.v043.i08

Gasparrini, A., Armstrong, B. and Kenward, M. G. (2010). Distributed lag non-linear models. *Stat. Med.* **29** 2224–2234. MR2757147 https://doi.org/10.1002/sim.3940

Gasparrini, A., Scheipl, F., Armstrong, B. and Kenward, M. G. (2017). A penalized framework for distributed lag non-linear models. *Biometrics* **73** 938–948. MR3713127 https://doi.org/10.1111/biom.12645

Gibson, E. A., Nunez, Y., Abuawad, A., Zota, A. R., Renzetti, S., Devick, K. L., Gennings, C., Goldsmith, J., Coull, B. A. et al. (2019). An overview of methods to address distinct research questions on environmental mixtures: An application to persistent organic pollutants and leukocyte telomere length. *Environ. Health* **18** 76. https://doi.org/10.1186/s12940-019-0515-1

Hamra, G. B. and Buckley, J. P. (2018). Environmental exposure mixtures: Questions and methods to address them. *Current Epidemiology Reports* **5** 160–165. https://doi.org/10.1007/s40471-018-0145-0

Heaton, M. J. and Peng, R. D. (2012). Flexible distributed lag models using random functions with application to estimating mortality displacement from heat-related deaths. *J. Agric. Biol. Environ. Stat.* **17** 313–331. MR2993269 https://doi.org/10.1007/s13253-012-0097-7

Herring, A. H. (2010). Nonparametric Bayes shrinkage for assessing exposures to mixtures subject to limits of detection. *Epidemiology* **21** S71–S76. https://doi.org/10.1097/EDE.0b013e3181cf0058

Hsu, H.-H. L., Chiu, Y.-H. M., Coull, B. A., Kloog, I., Schwartz, J., Lee, A., Wright, R. O. and Wright, R. J. (2015). Prenatal particulate air pollution and asthma onset in urban children. Identifying sensitive windows and sex differences. *Am. J. Respir. Crit. Care Med.* **192** 1052–1059. https://doi.org/10.1164/rccm.201504-0658OC

Keil, A. P., Buckley, J. P., O'Brien, K. M., Ferguson, K. K., Zhao, S. and White, A. J. (2020). A quantile-based g-computation approach to addressing the effects of exposure mixtures. *Environ. Health Perspect.* **128** 047004. https://doi.org/10.1289/EHP5838

LAKSHMANAN, A., CHIU, Y.-H. M., COULL, B. A., JUST, A. C., MAXWELL, S. L., SCHWARTZ, J., GRYPARIS, A., KLOOG, I., WRIGHT, R. J. et al. (2015). Associations between prenatal traffic-related air pollution exposure and birth weight: Modification by sex and maternal pre-pregnancy body mass index. *Environ. Res.* **137** 268–277. https://doi.org/10.1016/j.envres.2014.10.035

LEE, A., HSU, H.-H. L., CHIU, Y.-H. M., BOSE, S., ROSA, M. J., KLOOG, I., WILSON, A., SCHWARTZ, J., COHEN, S. et al. (2018). Prenatal fine particulate exposure and early childhood asthma: Effect of maternal stress and fetal sex. *J. Allergy Clin. Immunol.* **141** 1880–1886. https://doi.org/10.1016/j.jaci.2017.07.017

LIU, D., LIN, X. and GHOSH, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* **63** 1079–1088, 1311. MR2414585 https://doi.org/10.1111/j.1541-0420.2007.00799.x

LIU, S. H., BOBB, J. F., LEE, K. H. et al. (2018). Lagged kernel machine regression for identifying time windows of susceptibility to exposures of complex mixtures. *Biostatistics* **19** 325–341. MR3815175 https://doi.org/10.1093/biostatistics/kxx036

MOLITOR, J., PAPATHOMAS, M., JERRETT, M. and RICHARDSON, S. (2010). Bayesian profile regression with an application to the national survey of children's health. *Biostatistics* **11** 484–498. https://doi.org/10.1093/biostatistics/kxq013

MORK, D. and WILSON, A. (2021). Treed distributed lag nonlinear models. *Biostatistics*. https://doi.org/10.1093/biostatistics/kxaa051

MORRIS, J. S. (2015). Functional regression. *Annu. Rev. Stat. Appl.* **2** 321–359. https://doi.org/10.1146/annurev-statistics-010814-020413

MURRAY, I., ADAMS, R. P. and MACKAY, D. J. C. (2009). Elliptical slice sampling. *J. Mach. Learn. Res. Workshop Conf. Proc.* **9** 541–548.

PARK, S. K., TAO, Y., MEEKER, J. D., HARLOW, S. D. and MUKHERJEE, B. (2014). Environmental risk score as a new tool to examine multi-pollutants in epidemiologic research: An example from the NHANES study using serum lipid levels. *PLoS ONE* **9** e98632. https://doi.org/10.1371/journal.pone.0098632

PEARCE, J. L., WALLER, L. A., CHANG, H. H., KLEIN, M., MULHOLLAND, J. A., SARNAT, J. A., SARNAT, S. E., STRICKLAND, M. J. and TOLBERT, P. E. (2014). Using self-organizing maps to develop ambient air quality classifications: A time series example. *Environ. Health* **13** 56. https://doi.org/10.1186/1476-069X-13-56

PENG, R. D., DOMINICI, F. and WELTY, L. J. (2009). A Bayesian hierarchical distributed lag model for estimating the time course of risk of hospitalization associated with particulate matter air pollution. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 3–24. MR2662231 https://doi.org/10.1111/j.1467-9876.2008.00640.x

TAYLOR, K. W., JOUBERT, B. R., BRAUN, J. M., DILWORTH, C., GENNINGS, C., HAUSER, R., HEINDEL, J. J., RIDER, C. V., WEBSTER, T. F. et al. (2016). Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: Lessons from an innovative workshop. *Environ. Health Perspect.* **124** 227–229. https://doi.org/10.1289/EHP547

WARREN, J., FUENTES, M., HERRING, A. and LANGLOIS, P. (2012). Spatial-temporal modeling of the association between air pollution exposure and preterm birth: Identifying critical windows of exposure. *Biometrics* **68** 1157–1167. MR3040022 https://doi.org/10.1111/j.1541-0420.2012.01774.x

WARREN, J. L., FUENTES, M., HERRING, A. H. and LANGLOIS, P. H. (2013). Air pollution metric analysis while determining susceptible periods of pregnancy for low birth weight. *ISRN Obstetrics and Gynecology* **2013** 1–9. https://doi.org/10.1155/2013/387452

WARREN, J. L., STINGONE, J. A., HERRING, A. H. et al. (2016). Bayesian multinomial probit modeling of daily windows of susceptibility for maternal $PM_{2.5}$ exposure and congenital heart defects. *Stat. Med.* **35** 2786–2801. MR3513718 https://doi.org/10.1002/sim.6891

WARREN, J. L., KONG, W., LUBEN, T. J. and CHANG, H. H. (2020). Critical window variable selection: Estimating the impact of air pollution on very preterm birth. *Biostatistics* **21** 790–806. MR4164058 https://doi.org/10.1093/biostatistics/kxz006

WILSON, A., CHIU, Y.-H. M., HSU, H.-H. L., WRIGHT, R. O., WRIGHT, R. J. and COULL, B. A. (2017a). Potential for bias when estimating critical windows for air pollution in children's health. *Am. J. Epidemiol.* **186** 1281–1289. https://doi.org/10.1093/aje/kwx184

WILSON, A., CHIU, Y.-H. M., HSU, H.-H. L., WRIGHT, R. O., WRIGHT, R. J. and COULL, B. A. (2017b). Bayesian distributed lag interaction models to identify perinatal windows of vulnerability in children's health. *Biostatistics* **18** 537–552. MR3799593 https://doi.org/10.1093/biostatistics/kxx002

WILSON, A., HSU, H.-H. L., CHIU, Y.-H. M., WRIGHT, R. O., WRIGHT, R. J. and COULL, B. A. (2022a). Supplement to "Kernel machine and distributed lag models for assessing windows of susceptibility to environmental mixtures in children's health studies." https://doi.org/10.1214/21-AOAS1533SUPPA

WILSON, A., HSU, H.-H. L., CHIU, Y.-H. M., WRIGHT, R. O., WRIGHT, R. J. and COULL, B. A. (2022b). Supplement to "Kernel machine and distributed lag models for assessing windows of susceptibility to environmental mixtures in children's health studies." https://doi.org/10.1214/21-AOAS1533SUPPB

WOODRUFF, T. J., ZOTA, A. R. and SCHWARTZ, J. M. (2011). Environmental chemicals in pregnant women in the United States: NHANES 2003–2004. *Environ. Health Perspect.* **119** 878–885. https://doi.org/10.1289/ehp.1002727

WRIGHT, R. O. (2017). Environment, susceptibility windows, development, and child health. *Curr. Opin. Pediatr.* **29** 211–217. https://doi.org/10.1097/MOP.0000000000000465

WRIGHT, R. J., SUGLIA, S. F., LEVY, J., FORTUN, K., SHIELDS, A., SUBRAMANIAN, S. and WRIGHT, R. (2008). Transdisciplinary research strategies for understanding socially patterned disease: The asthma coalition on community, environment, and social stress (ACCESS) project as a case study. *Ciênc. Saúde Colet.* **13** 1729–1742. https://doi.org/10.1590/S1413-81232008000600008

XIA, Y. (2008). A multiple-index model and dimension reduction. *J. Amer. Statist. Assoc.* **103** 1631–1640. MR2504209 https://doi.org/10.1198/016214508000000805

ZANOBETTI, A., WAND, M. P., SCHWARTZ, J. and RYAN, L. M. (2000). Generalized additive distributed lag models: Quantifying mortality displacement. *Biostatistics* **1** 279–92. https://doi.org/10.1093/biostatistics/1.3.279

ZANOBETTI, A., AUSTIN, E., COULL, B. A., SCHWARTZ, J. and KOUTRAKIS, P. (2014). Health effects of multi-pollutant profiles. *Environ. Int.* **71** 13–19. https://doi.org/10.1016/j.envint.2014.05.023

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1111–1129
https://doi.org/10.1214/21-AOAS1535

DETECTING HETEROGENEOUS TREATMENT EFFECTS WITH INSTRUMENTAL VARIABLES AND APPLICATION TO THE OREGON HEALTH INSURANCE EXPERIMENT

BY MICHAEL JOHNSON[1,a], JIONGYI CAO[2,c] AND HYUNSEUNG KANG[1,b]

[1]*Department of Statistics, University of Wisconsin-Madison,* [a]*mwjohnson8@wisc.edu,* [b]*hyunseung@stat.wisc.edu*
[2]*Department of Statistics, University of Chicago,* [c]*jiongyi@uchicago.edu*

There is an increasing interest in estimating heterogeneity in causal effects in randomized and observational studies. However, little research has been conducted to understand effect heterogeneity in an instrumental variables study. In this work we present a method to estimate heterogeneous causal effects using an instrumental variable with matching. The method has two parts. The first part uses subject-matter knowledge and interpretable machine-learning techniques, such as classification and regression trees, to discover potential effect modifiers. The second part uses closed testing to test for statistical significance of each effect modifier while strongly controlling the familywise error rate. We apply this method on the Oregon Health Insurance Experiment, estimating the effect of Medicaid on the number of days an individual's health does not impede their usual activities by using a randomized lottery as an instrument. Our method revealed Medicaid's effect was most impactful among older, English-speaking, non-Asian males and younger, English-speaking individuals with, at most, a high school diploma or General Educational Development.

REFERENCES

ABADIE, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *J. Econometrics* **113** 231–263. MR1960380 https://doi.org/10.1016/S0304-4076(02)00201-4

AI, C. and CHEN, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* **71** 1795–1843. MR2015420 https://doi.org/10.1111/1468-0262.00470

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.

ATHEY, S. and IMBENS, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. Available at arXiv:1504.01132v1 [stat.ML].

ATHEY, S. and IMBENS, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA* **113** 7353–7360. MR3531135 https://doi.org/10.1073/pnas.1510489113

ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

BAIOCCHI, M., CHENG, J. and SMALL, D. S. (2014). Instrumental variable methods for causal inference. *Stat. Med.* **33** 2297–2340. MR3257582 https://doi.org/10.1002/sim.6128

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

BALKE, A. and PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *J. Amer. Statist. Assoc.* **92** 1171–1176.

BARGAGLI-STOFFI, F. J., DE-WITTE, K. and GNECCO, G. (2019). Heterogeneous causal effects with imperfect compliance: A novel Bayesian machine learning approach. Available at arXiv:1905.12707 [stat.ME].

BARGAGLI-STOFFI, F. J. and GNECCO, G. (2018). Estimating heterogeneous causal effects in the presence of irregular assignment mechanisms. In 2018 *IEEE 5th International Conference on Data Science and Advanced Analytics* (*DSAA*) 1–10. IEEE, New York.

Key words and phrases. Causal inference, complier average causal effect, heterogeneous treatment, instrumental variables, matching, machine learning, Oregon Health Insurance Experiment.

BLUNDELL, R., CHEN, X. and KRISTENSEN, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* **75** 1613–1669. MR2351452 https://doi.org/10.1111/j.1468-0262.2007.00808.x

BLUNDELL, R. and POWELL, J. L. (2003). Endogeneity in nonparametric and semiparametric regression models. *Econom. Soc. Monogr.* **36** 312–357.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. *Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

CHEN, X. and POUZO, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica* **80** 277–321. MR2920758 https://doi.org/10.3982/ECTA7888

CHERNOZHUKOV, V., DEMIRER, M., DUFLO, E. and FERNANDEZ-VAL, I. (2018). Generic machine learning inference on heterogenous treatment effects in randomized experiments. *National Bureau of Economic Research*.

DAROLLES, S., FAN, Y., FLORENS, J. P. and RENAULT, E. (2011). Nonparametric instrumental regression. *Econometrica* **79** 1541–1565. MR2883763 https://doi.org/10.3982/ECTA6539

DING, P. (2017). A paradox from randomization-based causal inference. *Statist. Sci.* **32** 331–345. MR3695995 https://doi.org/10.1214/16-STS571

FINKELSTEIN, A., TAUBMAN, S., WRIGHT, B., BERNSTEIN, M., GRUBER, J., NEWHOUSE, J. P., ALLEN, H., BAICKER, K. and GROUP, O. H. S. (2012). The Oregon health insurance experiment: Evidence from the first year. *Q. J. Econ.* **127** 1057–1106.

FOGARTY, C. B. (2018). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika* **105** 994–1000. MR3877880 https://doi.org/10.1093/biomet/asy034

FOGARTY, C. B. (2020). Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *J. Amer. Statist. Assoc.* **115** 1518–1530. MR4143482 https://doi.org/10.1080/01621459.2019.1632072

FOGARTY, C. B., LEE, K., KELZ, R. R. and KEELE, L. J. (2021). Biased encouragements and heterogeneous effects in an instrumental variable study of emergency general surgical outcomes. *J. Amer. Statist. Assoc.*

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

HALL, P. and HOROWITZ, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Ann. Statist.* **33** 2904–2929. MR2253107 https://doi.org/10.1214/009053605000000714

HERNÁN, M. A. and ROBINS, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology* **17** 360–372.

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

HODGES, J. L. JR. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Stat.* **34** 598–611. MR0152070 https://doi.org/10.1214/aoms/1177704172

HSU, J. Y., SMALL, D. S. and ROSENBAUM, P. R. (2013). Effect modification and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **108** 135–148. MR3174608 https://doi.org/10.1080/01621459.2012.742018

HSU, J. Y., ZUBIZARRETA, J. R., SMALL, D. S. and ROSENBAUM, P. R. (2015). Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika* **102** 767–782. MR3431552 https://doi.org/10.1093/biomet/asv034

IMBENS, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *J. Econ. Lit.* **48** 399–423.

JOHNSON, M., CAO, J. and KANG, H. (2022). Supplement to "Detecting heterogeneous treatment effects with instrumental variables and application to the Oregon health insurance experiment." https://doi.org/10.1214/21-AOAS1535SUPPA, https://doi.org/10.1214/21-AOAS1535SUPPB

KANG, H., PECK, L. and KEELE, L. (2018). Inference for instrumental variables: A randomization inference approach. *J. Roy. Statist. Soc. Ser. A* **181** 1231–1254. MR3876390 https://doi.org/10.1111/rssa.12353

KANG, H., KREUELS, B., ADJEI, O., KRUMKAMP, R., MAY, J. and SMALL, D. S. (2013). The causal effect of malaria on stunting: A Mendelian randomization and matching approach. *Int. J. Epidemiol.* **42** 1390–1398.

KANG, H., KREUELS, B., MAY, J. and SMALL, D. S. (2016). Full matching approach to instrumental variables estimation with application to the effect of malaria on stunting. *Ann. Appl. Stat.* **10** 335–364. MR3480499 https://doi.org/10.1214/15-AOAS894

LEE, K., BARGAGLI-STOFFI, F. J. and DOMINICI, F. (2021). Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. Preprint. Available at arXiv:2009.09036.

LEE, K., SMALL, D. S. and DOMINICI, F. (2021). Discovering heterogeneous exposure effects using randomization inference in air pollution studies. *J. Amer. Statist. Assoc.* **116** 569–580. MR4270004 https://doi.org/10.1080/01621459.2020.1870476

LEE, K., SMALL, D. S. and ROSENBAUM, P. R. (2018). A powerful approach to the study of moderate effect modification in observational studies. *Biometrics* **74** 1161–1170. MR3908134

LEE, K., SMALL, D. S., HSU, J. Y., SILBER, J. H. and ROSENBAUM, P. R. (2018). Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing. *J. Roy. Statist. Soc. Ser. A* **181** 535–546. MR3749529 https://doi.org/10.1111/rssa.12298

MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056 https://doi.org/10.1093/biomet/63.3.655

NEWEY, W. K. and POWELL, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71** 1565–1578. MR2000257 https://doi.org/10.1111/1468-0262.00459

PARK, C. and KANG, H. (2020). A groupwise approach for inferring heterogeneous treatment effects in causal inference. Preprint. Available at arXiv:1908.04427v2.

ROSENBAUM, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. MR1962487 https://doi.org/10.1214/ss/1042727942

ROSENBAUM, P. R. (2002b). [Covariance adjustment in randomized experiments and observational studies]: Rejoinder. *Statist. Sci.* **17** 321–327. With comments and a rejoinder by the author. MR1962487 https://doi.org/10.1214/ss/1042727942

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

ROSENBAUM, P. R. (2020). Modern algorithms for matching in observational studies. *Annu. Rev. Stat. Appl.* **7** 143–176. MR4104189 https://doi.org/10.1146/annurev-statistics-031219-041058

ROTHWELL, P. M. (2005). Subgroup analysis in randomised controlled trials: Importance, indications, and interpretation. *Lancet* **365** 176–186.

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

RUBIN, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* **2** 169–188.

STAIGER, D. and STOCK, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65** 557–586. MR1445622 https://doi.org/10.2307/2171753

STALLONES, R. A. (1987). The use and abuse of subgroup analysis in epidemiological research. *Prev. Med.* **16** 183–194.

STOCK, J. H., WRIGHT, J. H. and YOGO, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econom. Statist.* **20** 518–529. MR1973801 https://doi.org/10.1198/073500102288618658

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

SU, L., MURTAZASHVILI, I. and ULLAH, A. (2013). Local linear GMM estimation of functional coefficient IV models with an application to estimating the rate of return to schooling. *J. Bus. Econom. Statist.* **31** 184–207. MR3055331 https://doi.org/10.1080/07350015.2012.754314

SU, X., TSAI, C.-L., WANG, H., NICKERSON, D. M. and LI, B. (2009). Subgroup analysis via recursive partitioning. *J. Mach. Learn. Res.* **10** 141–158.

SWANSON, S. A. and HERNÁN, M. A. (2013). Commentary: How to report instrumental variable analyses (suggestions welcome). *Epidemiology* **24** 370–374.

SWANSON, S. A. and HERNÁN, M. A. (2014). Think globally, act globally: An epidemiologist's perspective on instrumental variable estimation [discussion of MR3264545]. *Statist. Sci.* **29** 371–374. MR3264549 https://doi.org/10.1214/14-STS491

THERNEAU, T., ATKINSON, B. and RIPLEY, B. (2015). Package 'rpart'. R package version 4.1-15. Available at https://cran.r-project.org/package=rpart.

WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839

WANG, T. and RUDIN, C. (2021). Causal rule sets for identifying subgroups with enhanced treatment effect. Preprint. Available at arXiv:1710.05426.

YU, R. (2019). bigmatch: Making optimal matching size-scalable using optimal calipers. R package version 0.6.1. Available at https://CRAN.R-project.org/package=bigmatch.

YUSUF, S., WITTES, J., PROBSTFIELD, J. and TYROLER, H. A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* **266** 93–98.

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1130–1150
https://doi.org/10.1214/21-AOAS1536

STATISTICAL SHAPE ANALYSIS OF BRAIN ARTERIAL NETWORKS (BAN)

BY XIAOYANG GUO[1,a], ADITI BASU BAL[1,b], TOM NEEDHAM[2,d] AND ANUJ SRIVASTAVA[1,c]

[1]*Department of Statistics, Florida State University,* [a]*xiaoyang.guo.fl@gmail.com,* [b]*ab18z@my.fsu.edu,* [c]*anuj@stat.fsu.edu*
[2]*Department of Mathematics, Florida State University,* [d]*tneedham@fsu.edu*

The arterial networks in the human brain, termed brain arterial networks or BANs, are complex arrangements of individual arteries, branching patterns, and interconnectivity. BANs play an essential role in characterizing and understanding brain physiology, and one would like tools for statistically analyzing the shapes of BANs. These tools include quantifying shape differences, comparing populations of subjects, and studying the effects of covariates on these shapes. This paper mathematically represents and statistically analyzes BAN shapes as *elastic shape graphs*. Each elastic shape graph consists of nodes, or points in 3D, connected by 3D curves, or edges, with arbitrary shapes. We develop a mathematical representation, a Riemannian metric and other geometrical tools, such as computations of geodesics, means, covariances, and PCA, for helping analyze BANs as elastic graphs. We apply this analysis to BANs after dividing them into four components—top, bottom, left, and right. The framework is then used to generate shape summaries of BANs from 92 subjects and study the effects of age and gender on shapes of BAN components. While gender effects require further investigation, we conclude that age has a clear, quantifiable effect on BAN shapes. Specifically, we find an increased variance in BAN shapes as age increases.

REFERENCES

AYDIN, B., PATAKI, G., WANG, H., BULLITT, E. and MARRON, J. S. (2009). A principal component analysis for trees. *Ann. Appl. Stat.* **3** 1597–1615. MR2752149 https://doi.org/10.1214/09-AOAS263

AYLWARD, S. R. and BULLITT, E. (2002). Initialization, noise, singularities, and scale in height ridge traversal for tubular object centerline extraction. *IEEE Trans. Med. Imag.* **21** 61–75. https://doi.org/10.1109/42.993126

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

BIGOT, J., GOUET, R., KLEIN, T. and LÓPEZ, A. (2017). Geodesic PCA in the Wasserstein space by convex PCA. *Ann. Inst. Henri Poincaré Probab. Stat.* **53** 1–26. MR3606732 https://doi.org/10.1214/15-AIHP706

BULLITT, E., ZENG, D., GERIG, G., AYLWARD, S., JOSHI, S., SMITH, J. K., LIN, W. and EWEND, M. G. (2005). Vessel tortuosity and brain tumor malignancy: A blinded study1. *Acad. Radiol.* **12** 1232–1240.

BULLITT, E., ZENG, D., MORTAMET, B., GHOSH, A., AYLWARD, S. R., LIN, W., MARKS, B. L. and SMITH, K. (2010). The effects of healthy aging on intracerebral blood vessels visualized by magnetic resonance angiography. *Neurobiol. Aging* **31** 290–300.

CAELLI, T. and KOSINOV, S. (2004). An eigenspace projection clustering method for inexact graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 515–519.

CALISSANO, A., FERAGEN, A. and VANTINI, S. (2020). Populations of unlabeled networks: Graph space geometry and geodesic principal components.

CAZELLES, E., SEGUY, V., BIGOT, J., CUTURI, M. and PAPADAKIS, N. (2018). Geodesic PCA versus log-PCA of histograms in the Wasserstein space. *SIAM J. Sci. Comput.* **40** B429–B456. MR3780753 https://doi.org/10.1137/17M1143459

CHAKRABORTY, R. and VEMURI, B. C. (2019). Statistics on the Stiefel manifold: Theory and applications. *Ann. Statist.* **47** 415–438. MR3909937 https://doi.org/10.1214/18-AOS1692

CHOWDHURY, S. and NEEDHAM, T. (2020). Gromov-Wasserstein averaging in a Riemannian framework. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*) *Workshops*.

COUR, T., SRINIVASAN, P. and SHI, J. (2007). Balanced graph matching. In *Advances in Neural Information Processing Systems* 313–320.

Key words and phrases. Statistical shape analysis, graph matching, brain artery network, graph shape PCA.

DRYDEN, I. L. and MARDIA, K. V. (2016). *Statistical Shape Analysis with Applications in R*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3559734 https://doi.org/10.1002/9781119072492

DUNCAN, A., KLASSEN, E. and SRIVASTAVA, A. (2018). Statistical shape analysis of simplified neuronal trees. *Ann. Appl. Stat.* **12** 1385–1421. MR3852682 https://doi.org/10.1214/17-AOAS1107

FISHKIND, D. E., ADALI, S., PATSOLIC, H. G., MENG, L., SINGH, D., LYZINSKI, V. and PRIEBE, C. E. (2012). Seeded graph matching.

GOLD, S. and RANGARAJAN, A. (1996). A graduated assignment algorithm for graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **18** 377–388.

GUO, X., SRIVASTAVA, A. and SARKAR, S. (2021). A quotient space formulation for generative statistical analysis of graphical data. *J. Math. Imaging Vision* **63** 735–752. MR4271992 https://doi.org/10.1007/s10851-021-01027-1

GUO, X., BASU BAL, A., NEEDHAM, T. and SRIVASTAVA, A. (2022). Supplement to "Statistical shape analysis of brain arterial networks (BAN)." https://doi.org/10.1214/21-AOAS1536SUPP

HAGWOOD, C., BERNAL, J., HALTER, M., ELLIOTT, J. and BRENNAN, T. (2013). Testing equality of cell populations based on shape and geodesic distances. *IEEE Trans. Med. Imag.* **32**.

HOOVER, A., KOUZNETSOVA, V. and GOLDBAUM, M. (2000). Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Trans. Med. Imag.* **19** 203–210.

HUCKEMANN, S., HOTZ, T. and MUNK, A. (2010). Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statist. Sinica* **20** 1–58. MR2640651

JAIN, B. J. and OBERMAYER, K. (2009). Structure spaces. *J. Mach. Learn. Res.* **10** 2667–2714. MR2576333

JAIN, B. J. and OBERMAYER, K. (2011). Graph quantization. *Comput. Vis. Image Underst.* **115** 946–961.

JAIN, B. J. and OBERMAYER, K. (2012). Learning in Riemannian orbifolds.

JERMYN, I. H., KURTEK, S., KLASSEN, E. and SRIVASTAVA, A. (2012). Elastic shape matching of parameterized surfaces using square root normal fields. In *European Conference on Computer Vision* 804–817. Springer, Berlin.

JERMYN, I. H., KURTEK, S., LAGA, H. and SRIVASTAVA, A. (2017). Elastic shape analysis of three-dimensional objects. *Synthesis Lectures on Computer Vision* **12** 1–185.

KENDALL, D. G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *Bull. Lond. Math. Soc.* **16** 81–121. MR0737237 https://doi.org/10.1112/blms/16.2.81

KENDALL, D. G., BARDEN, D., CARNE, T. K. and LE, H. (1999). *Shape and Shape Theory. Wiley Series in Probability and Statistics*. Wiley, Chichester. MR1891212 https://doi.org/10.1002/9780470317006

KLASSEN, E., SRIVASTAVA, A., MIO, M. and JOSHI, S. H. (2004). Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 372–383.

KONG, J.-H., FISH, D. R., ROCKHILL, R. L. and MASLAND, R. H. (2005). Diversity of ganglion cells in the mouse retina: Unsupervised morphological classification and its limits. *J. Comp. Neurol.* **489** 293–310.

KOOPMANS, T. C. and BECKMANN, M. (1957). Assignment problems and the location of economic activities. *Econometrica* **25** 53–76. MR0089106 https://doi.org/10.2307/1907742

LAWLER, E. L. (1962/63). The quadratic assignment problem. *Manage. Sci.* **9** 586–599. MR0152361 https://doi.org/10.1287/mnsc.9.4.586

LE BRIGANT, A. (2019). A discrete framework to find the optimal matching between manifold-valued curves. *J. Math. Imaging Vision* **61** 40–70. MR3900062 https://doi.org/10.1007/s10851-018-0820-2

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 Press, 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.

LIU, Z.-Y., QIAO, H. and XU, L. (2012). An extended path following algorithm for graph-matching problem. *IEEE Trans. Pattern Anal. Mach. Intell.* **34** 1451–1456.

SHEN, D., SHEN, H., BHAMIDI, S., MUÑOZ MALDONADO, Y., KIM, Y. and MARRON, J. S. (2014). Functional data analysis of tree data objects. *J. Comput. Graph. Statist.* **23** 418–438. MR3215818 https://doi.org/10.1080/10618600.2013.786943

SMALL, C. G. (1996). *The Statistical Theory of Shape. Springer Series in Statistics*. Springer, New York. MR1418639 https://doi.org/10.1007/978-1-4612-4032-7

SONNENSCHEIN, A., VANDERZEE, D., PITCHERS, W. R., CHARI, S. and DWORKIN, I. (2015). An image database of Drosophila melanogaster wings for phenomic and biometric analysis. *GigaScience* **4** 25. https://doi.org/10.1186/s13742-015-0065-6

SRIVASTAVA, A. and KLASSEN, E. P. (2016). *Functional and Shape Data Analysis. Springer Series in Statistics*. Springer, New York. MR3821566

SRIVASTAVA, A., KLASSEN, E., JOSHI, S. H. and JERMYN, I. H. (2011). Shape analysis of elastic curves in Euclidean spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 1415–1428.

UMEYAMA, S. (1988). An eigendecomposition approach to weighted graph matching problems. *IEEE Trans. Pattern Anal. Mach. Intell.* **10** 695–703.

VOGELSTEIN, J. T., CONROY, J. M., LYZINSKI, V., PODRAZIK, L. J., KRATZER, S. G., HARLEY, E. T., FISHKIND, D. E., VOGELSTEIN, R. J. and PRIEBE, C. E. (2015). Fast approximate quadratic programming for graph matching. *PLoS ONE* **10** e0121002.

WANG, G., LAGA, H., JIA, J., MIKLAVCIC, S. J. and SRIVASTAVA, A. (2020). Statistical analysis and modeling of the geometry and topology of plant roots. *J. Theoret. Biol.* **486** 110108, 11. MR4042643 https://doi.org/10.1016/j.jtbi.2019.110108

WASSERMAN, L. (2018). Topological data analysis. *Annu. Rev. Stat. Appl.* **5** 501–535. MR3774757 https://doi.org/10.1146/annurev-statistics-031017-100045

YAN, J., YIN, X.-C., LIN, W., DENG, C., ZHA, H. and YANG, X. (2016). A short survey of recent advances in graph matching. In *Proceedings of the* 2016 *ACM on International Conference on Multimedia Retrieval* 167–174.

YOUNES, L. (1998). Computable elastic distances between shapes. *SIAM J. Appl. Math.* **58** 565–586. MR1617630 https://doi.org/10.1137/S0036139995287685

YOUNES, L., MICHOR, P. W., SHAH, J. and MUMFORD, D. (2008). A metric on shape space with explicit geodesics. *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei* (9) *Mat. Appl.* **19** 25–57. MR2383560 https://doi.org/10.4171/RLM/506

ZANFIR, A. and SMINCHISESCU, C. (2018). Deep learning of graph matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2684–2693.

ZHANG, Z., KLASSEN, E. and SRIVASTAVA, A. (2018). Phase-amplitude separation and modeling of spherical trajectories. *J. Comput. Graph. Statist.* **27** 85–97. MR3788303 https://doi.org/10.1080/10618600.2017.1340892

ZHOU, F. and DE LA TORRE, F. (2012). Factorized graph matching. In 2012 *IEEE Conference on Computer Vision and Pattern Recognition* 127–134. IEEE, New York.

ZHOU, F. and DE LA TORRE, F. (2016). Factorized graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **38** 1774–1789. https://doi.org/10.1109/TPAMI.2015.2501802

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1151–1170
https://doi.org/10.1214/21-AOAS1538

SPATIOTEMPORAL-TEXTUAL POINT PROCESSES FOR CRIME LINKAGE DETECTION

BY SHIXIANG ZHU[a] AND YAO XIE[b]

School of Industrial and Systems Engineering, Georgia Institute of Technology, [a]*shixiang.zhu@gatech.edu,* [b]*Yao.xie@isye.gatech.edu*

Crimes emerge out of complex interactions of human behaviors and situations. Linkages between crime incidents are highly complex. Detecting crime linkage, given a set of incidents, is a highly challenging task since we only have limited information, including text descriptions, incident times, and locations. In practice, there are very few labels. We propose a new statistical modeling framework for *spatiotemporal-textual* data and demonstrate its usage on crime linkage detection. We capture linkages of crime incidents via multivariate marked spatiotemporal Hawkes processes and treat embedding vectors of the free-text as *marks* of the incident, inspired by the notion of modus operandi (M.O.) in crime analysis. Numerical results, using real data, demonstrate the good performance of our method as well as reveals interesting patterns in the crime data: the joint modeling of space, time, and text information enhances crime linkage detection, compared with the state-of-the-art, and the learned spatial dependence from data can be useful for police operations.

REFERENCES

ADDERLEY, R. (2004). The use of data mining techniques in operational crime fighting. In *Intelligence and Security Informatics* 418–425. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-25952-7_32

ADDERLEY, R. and MUSGROVE, P. (2003). Modus operandi modelling of group offending: A data-mining case study. *Int. J. Police Sci. Manag.* **5** 265–276. https://doi.org/10.1350/ijps.5.4.265.24933

ANDRADE, D. C., ROCHA-JUNIOR, J. A. B. and COSTA, D. G. (2017). Efficient processing of spatio-temporal-textual queries. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*, *WebMedia'*17. 165–172. ACM, New York. https://doi.org/10.1145/3126858.3126877

BOUHANA, N. and JOHNSON, S. D. (2016). Consistency and specificity in burglars who commit prolific residential burglary: Testing the core assumptions underpinning behavioural crime linkage. *Legal Criminol. Psychol.* **21** 77–94. https://doi.org/10.1111/lcrp.12050

COCX, T. K. and KOSTERS, W. A. (2006). A distance measure for determining similarity between criminal investigations. In *Industrial Conference on Data Mining* 511–525. Springer, Berlin.

DAHBUR, K. and MUSCARELLO, T. (2003). Classification system for serial criminal patterns. *Artif. Intell. Law* **11** 251–269. https://doi.org/10.1023/B:ARTI.0000045994.96685.21

DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes. Vol. I: Elementary Theory and Methods*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR1950431

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

DU, N., DAI, H., TRIVEDI, R., UPADHYAY, U., GOMEZ-RODRIGUEZ, M. and SONG, L. (2016). Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *KDD'*16. 1555–1564. ACM, New York. https://doi.org/10.1145/2939672.2939875

FISCHER, A. and IGEL, C. (2012). An introduction to restricted Boltzmann machines. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Lecture Notes in Computer Science* **7441**. Springer, Heidelberg. https://doi.org/10.1007/978-3-642-33275-3_2

FOX, E. W., SCHOENBERG, F. P. and GORDON, J. S. (2016). Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *Ann. Appl. Stat.* **10** 1725–1756. MR3553242 https://doi.org/10.1214/16-AOAS957

Key words and phrases. Hawkes processes with marks, crime linkage detection, variable selection.

GOMAA, W. H. and FAHMY, A. A. (2013). Article: A survey of text similarity approaches. *Int. J. Comput. Appl.* **68** 13–18.

HALKIAS, X., PARIS, S. and GLOTIN, H. (2013). Sparse penalty in deep belief networks: Using the mixed norm constraint.

HARRIS, Z. S. (1954). Distributional structure. *Word* **10** 146–162. https://doi.org/10.1080/00437956.1954.11659520

HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. MR0278410 https://doi.org/10.1093/biomet/58.1.83

HINTON, G. E. (2002). Training products of experts by minimizing constractive divergence. *Neural Comput.* **14** 1771–1800. https://doi.org/10.1162/089976602760128018

HINTON, G. E. (2012). A practical guide to training restricted Boltzmann machines. In *Neural Networks*: *Tricks of the Trade* 599–619. Springer, Berlin.

HONG, S., WU, M., LI, H. and WU, Z. (2017). Event2vec: Learning representations of events on temporal sequences. In *Web and Big Data* 33–47. Springer, Berlin.

KEYVANRAD, M. A. and HOMAYOUNPOUR, M. M. (2017). Effective sparsity control in deep belief networks using normal regularization term. *Knowl. Inf. Syst.* **53** 533–550. https://doi.org/10.1007/s10115-017-1049-x

KUANG, D., BRANTINGHAM, P. J. and BERTOZZI, A. L. (2017). Crime topic modeling. *Crime Science* **6** 12. https://doi.org/10.1186/s40163-017-0074-0

LAI, E. L., MOYER, D., YUAN, B., FOX, E., HUNTER, B., BERTOZZI, A. L. and BRANTINGHAM, P. J. (2016). Topic time series analysis of microblogs. *IMA J. Appl. Math.* **81** 409–431. MR3564661 https://doi.org/10.1093/imamat/hxw025

LAN, G. (2020). *First-Order and Stochastic Optimization Methods for Machine Learning*. *Springer Series in the Data Sciences*. Springer, Cham. MR4219819 https://doi.org/10.1007/978-3-030-39568-1

LIN, S. and BROWN, D. E. (2006). An outlier-based data association method for linking criminal incidents. *Legal Criminol. Psychol.* **41** 604–615. https://doi.org/10.1016/j.dss.2004.06.005

LIU, X., JIAN, C. and LU, C.-T. (2010). A spatio-temporal-textual crime search engine. In *Proceedings of the* 18*th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, *GIS*'10. 528–529. ACM, New York. https://doi.org/10.1145/1869790.1869881

LUO, H., SHEN, R., NIU, C. and ULLRICH, C. (2011). Sparse group restricted Boltzmann machines. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, *AAAI*'11. 429–434. AAAI Press, Menlo Park.

MA, L., CHEN, Y. and HUANG, H. (2010). AK-modes: A weighted clustering algorithm for finding similar case subsets. In 2010 *IEEE International Conference on Intelligent Systems and Knowledge Engineering* 218–223. https://doi.org/10.1109/ISKE.2010.5680876

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

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. and DEAN, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the* 26*th International Conference on Neural Information Processing Systems—Volume* 2, *NIPS*'13. 3111–3119. Curran Associates, Red Hook.

MOHLER, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast.* **30** 491–497.

MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. MR2816705 https://doi.org/10.1198/jasa.2011.ap09546

NATH, S. V. (2006). Crime pattern detection using data mining. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, IEEE Press, New York. https://doi.org/10.1109/WI-IATW.2006.55

PARK, J., SCHOENBERG, F. P., BERTOZZI, A. L. and BRANTINGHAM, P. J. (2021). Investigating clustering and violence interruption in gang-related violent crime data using spatial–temporal point processes with covariates. *J. Amer. Statist. Assoc.* 1–14.

PORTER, M. D. (2016). A statistical approach to crime linkage. *Amer. Statist.* **70** 152–165. MR3511045 https://doi.org/10.1080/00031305.2015.1123185

QUINN, C. J., COLEMAN, T. P., KIYAVASH, N. and HATSOPOULOS, N. G. (2011). Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *J. Comput. Neurosci.* **30** 17–44. MR2774388 https://doi.org/10.1007/s10827-010-0247-2

RANZATO, M. A., BOUREAU, Y.-L. and LECUN, Y. (2007). Sparse feature learning for deep belief networks. In *Proceedings of the* 20*th International Conference on Neural Information Processing Systems*, *NIPS*'07. 1185–1192. Curran Associates, Red Hook.

RANZATO, M., POULTNEY, C., CHOPRA, S. and LECUN, Y. (2006). Efficient learning of sparse representations with an energy-based model. In *Proceedings of the* 19*th International Conference on Neural Information Processing Systems*, *NIPS'*06. 1137–1144. MIT Press, Cambridge.

RASMUSSEN, J. G. (2011). Temporal point processes: The conditional intensity function.

REINHART, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statist. Sci.* **33** 299–318. MR3843374 https://doi.org/10.1214/17-STS629

ROUSSEEUW, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20** 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

SHEN, T., JIANG, J., LIN, W., GE, J., WU, P., ZHOU, Y., ZUO, C., WANG, J., YAN, Z. et al. (2019). Use of overlapping group LASSO sparse deep belief network to discriminate Parkinson's disease and normal control. *Front. Neurosci.* **13** 396. https://doi.org/10.3389/fnins.2019.00396

SIMMA, A. and JORDAN, M. I. (2010). Modeling events with cascades of Poisson processes. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, *UAI'*10. 546–555. AUAI Press, Catalina Island, CA.

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

VEEN, A. and SCHOENBERG, F. P. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm. *J. Amer. Statist. Assoc.* **103** 614–624. MR2523998 https://doi.org/10.1198/016214508000000148

WANG, B., DONG, H., BOEDIHARDJO, A. P., LU, C.-T., YU, H., CHEN, I.-R. and DAI, J. (2012). An integrated framework for spatio-temporal-textual search and mining. In *Proceedings of the* 20*th International Conference on Advances in Geographic Information Systems* 570–573.

WANG, T., RUDIN, C., WAGNER, D. and SEVIERI, R. (2015). Finding patterns with a rotten core: Data mining for crime series with cores. *Big Data* **3** 3–21. https://doi.org/10.1089/big.2014.0021

WOODHAMS, J., BULL, R. and HOLLIN, C. R. (2007). *Case Linkage*. Springer, Berlin. https://doi.org/10.1007/978-1-60327-146-2_6

XU, H., FARAJTABAR, M. and ZHA, H. (2016). Learning granger causality for Hawkes processes. In *Proceedings of the* 33*rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.). *Proceedings of Machine Learning Research* **48** 1717–1726. PMLR, New York, New York, USA.

ZHANG, C., LIU, L., LEI, D., YUAN, Q., ZHUANG, H., HANRATTY, T. and HAN, J. (2017). TrioVecEvent: Embedding-based online local event detection in geo-tagged tweet streams. In *ACM SIGKDD International Conference* 595–604. https://doi.org/10.1145/3097983.3098027

ZHU, S. and XIE, Y. (2018). Crime incidents embedding using restricted Boltzmann machines. In 2018 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*) 2376–2380. https://doi.org/10.1109/ICASSP.2018.8461621

ZHU, S. and XIE, Y. (2019). Crime event embedding with unsupervised feature selection. In *ICASSP* 2019—2019 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*) 3922–3926. https://doi.org/10.1109/ICASSP.2019.8682285

ZHU, S. and XIE, Y. (2022). Supplement to "SpatioTemporal-Textual Point Processes for Crime Linkage Detection." https://doi.org/10.1214/21-AOAS1538SUPPA, https://doi.org/10.1214/21-AOAS1538SUPPB, https://doi.org/10.1214/21-AOAS1538SUPPC

ZHUANG, J., OGATA, Y. and VERE-JONES, D. (2002). Stochastic declustering of space-time earthquake occurrences. *J. Amer. Statist. Assoc.* **97** 369–380. MR1941459 https://doi.org/10.1198/016214502760046925

ZHUANG, J., OGATA, Y. and VERE-JONES, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *J. Geophys. Res.*, *Solid Earth* **109**.

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1171–1190
https://doi.org/10.1214/21-AOAS1539

MARKOV-MODULATED HAWKES PROCESSES FOR MODELING SPORADIC AND BURSTY EVENT OCCURRENCES IN SOCIAL INTERACTIONS

BY JING WU[1,a] OWEN G. WARD[1,b], JAMES CURLEY[2,d] AND TIAN ZHENG[1,c]

[1]*Department of Statistics, Columbia University,* [a]*ju3233@columbia.edu,* [b]*owen.ward@columbia.edu,* [c]*tian.zheng@columbia.edu*

[2]*Department of Psychology, University of Texas at Austin,* [d]*curley@utexas.edu*

Modeling event dynamics is central to many disciplines. Patterns in observed social interaction events can be commonly modeled using point processes. Such social interaction event data often exhibit self-exciting, heterogeneous and sporadic trends which is challenging for conventional models. It is reasonable to assume that there exists a hidden state process that drives different event dynamics at different states. In this paper we propose a Markov modulated Hawkes process (MMHP) model for learning such a mixture of social interaction event dynamics and develop corresponding inference algorithms. Numerical experiments using synthetic data demonstrate that MMHP with the proposed estimation algorithms consistently recover the true hidden state process in simulations, while email data from a large university and data from an animal behavior study show that the procedure captures distinct event dynamics that reveal interesting social structures in the real data.

REFERENCES

BARABASI, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature* **435** 207–211.

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

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

BROWN, E. N., BARBIERI, R., VENTURA, V., KASS, R. E. and FRANK, L. M. (2002). The time-rescaling theorem and its application to neural spike train data analysis. *Neural Comput.* **14** 325–346.

COHN, I., EL-HAY, T., FRIEDMAN, N. and KUPFERMAN, R. (2009). Mean field variational approximation for continuous-time Bayesian networks. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* 91–100. AUAI Press.

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

DU, N., FARAJTABAR, M., AHMED, A., SMOLA, A. J. and SONG, L. (2015). Dirichlet–Hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 219–228.

FISCHER, W. and MEIER-HELLSTERN, K. (1993). The Markov-modulated Poisson process (MMPP) cookbook. *Perform. Eval.* **18** 149–171. MR1237373 https://doi.org/10.1016/0166-5316(93)90035-S

FORNEY, G. D. JR. (1973). The Viterbi algorithm. *Proc. IEEE* **61** 268–278. MR0439384

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

GUO, J., BETANCOURT, M., BRUBAKER, M., CARPENTER, B., GOODRICH, B., HOFFMAN, M., LEE, D., MALECKI, M. and GELMAN, A. (2014). RStan: The R interface to Stan.

HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. MR0278410 https://doi.org/10.1093/biomet/58.1.83

HAWKES, A. G. (2018). Hawkes processes and their applications to finance: A review. *Quant. Finance* **18** 193–198. MR3750729 https://doi.org/10.1080/14697688.2017.1403131

Key words and phrases. Social interaction dynamics, self-exciting processes, heterogeneous point processes, latent Markov processes, Bayesian inference.

KOSSINETS, G. and WATTS, D. J. (2006). Empirical analysis of an evolving social network. *Science* **311** 88–90. MR2192483 https://doi.org/10.1126/science.1116869

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

LEIVA, D., SOLANAS, A. and SALAFRANCA, L. (2008). Testing reciprocity in social interactions: A comparison between the directional consistency and skew-symmetry statistics. *Behav. Res. Methods* **40** 626–634.

LINDERMAN, S. and ADAMS, R. (2014). Discovering latent network structure in point process data. In *International Conference on Machine Learning* 1413–1421.

LIU, Y., GELMAN, A. and ZHENG, T. (2015). Simulation-efficient shortest probability intervals. *Stat. Comput.* **25** 809–819. MR3360494 https://doi.org/10.1007/s11222-015-9563-8

MCDONALD, D. B. and SHIZUKA, D. (2012). Comparative transitive and temporal orderliness in dominance networks. *Behav. Ecol.* **24** 511–520.

OGATA, Y. (1981). On Lewis' simulation method for point processes. *IEEE Trans. Inf. Theory* **27** 23–31.

PERRY, P. O. and WOLFE, P. J. (2013). Point process modelling for directed interaction networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 821–849. MR3124793 https://doi.org/10.1111/rssb.12013

RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77** 257–286.

RAO, V. and TEH, Y. W. (2013). Fast MCMC sampling for Markov jump processes and extensions. *J. Mach. Learn. Res.* **14** 3295–3320. MR3144463

SCOTT, S. L. and SMYTH, P. (2003). The Markov modulated Poisson process and Markov Poisson cascade with applications to web traffic modeling. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) 671–680. Oxford Univ. Press, New York. MR2003531

SIMMA, A. and JORDAN, M. I. (2010). Modeling events with cascades of Poisson processes. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* 546–555.

SO, N., FRANKS, B., LIM, S. and CURLEY, J. P. (2015). A social network approach reveals associations between mouse social dominance and brain gene expression. *PLoS ONE* **10** e0134509. https://doi.org/10.1371/journal.pone.0134509

VEEN, A. and SCHOENBERG, F. P. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm. *J. Amer. Statist. Assoc.* **103** 614–624. MR2523998 https://doi.org/10.1198/016214508000000148

VRIES, H. D. (1998). Finding a dominance order most consistent with a linear hierarchy: A new procedure and review. *Anim. Behav.* **55** 827–843.

WANG, T., BEBBINGTON, M. and HARTE, D. (2012). Markov-modulated Hawkes process with stepwise decay. *Ann. Inst. Statist. Math.* **64** 521–544. MR2880868 https://doi.org/10.1007/s10463-010-0320-7

WANG, Y., DU, N., TRIVEDI, R. and SONG, L. (2016). Coevolutionary latent feature processes for continuous-time user-item interactions. In *Advances in Neural Information Processing Systems* 4547–4555.

WEISS, J., NATARAJAN, S. and PAGE, D. (2012). Multiplicative forests for continuous-time processes. In *Advances in Neural Information Processing Systems* 458–466.

WILLIAMSON, C. M., LEE, W. and CURLEY, J. P. (2016). Temporal dynamics of social hierarchy formation and maintenance in male mice. *Anim. Behav.* **115** 259–272.

WILLIAMSON, C. M., ROMEO, R. D. and CURLEY, J. P. (2017). Dynamic changes in social dominance and mPOA GnRH expression in male mice following social opportunity. *Horm. Behav.* **87** 80–88. https://doi.org/10.1016/j.yhbeh.2016.11.001

WU, J., WARD, O. G., CURLEY, J. and ZHENG, T. (2022). Supplement to "Markov-modulated Hawkes processes for modeling sporadic and bursty event occurrences in social interactions." https://doi.org/10.1214/21-AOAS1539SUPP

XU, H. and ZHA, H. (2017). A Dirichlet mixture model of Hawkes processes for event sequence clustering. In *Advances in Neural Information Processing Systems* 1354–1363.

ZHAO, Q., ERDOGDU, M. A., HE, H. Y., RAJARAMAN, A. and LESKOVEC, J. (2015). Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1513–1522. ACM, New York.

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1191–1214
https://doi.org/10.1214/21-AOAS1542

CONDITIONAL FUNCTIONAL CLUSTERING FOR LONGITUDINAL DATA WITH HETEROGENEOUS NONLINEAR PATTERNS

BY TIANHAO WANG[a], LEI YU[b], SUE E. LEURGANS[c], ROBERT S. WILSON[d], DAVID A. BENNETT[e] AND PATRICIA A. BOYLE[f]

Rush Alzheimer's Disease Center, Rush University Medical Center, [a]*tianhao_wang@rush.edu,* [b]*lei_yu@rush.edu,* [c]*sue_e_leurgans@rush.edu,* [d]*robert_s_wilson@rush.edu,* [e]*david_a_bennett@rush.edu,* [f]*patricia_boyle@rush.edu*

In studies of cognitive aging, it is crucial to distinguish subtypes of longitudinal cognition change while accounting for the effects of given covariates. The longitudinal cognition trajectories and the covariate effects can both be nonlinear with heterogeneous shapes that do not follow a simple parametric form, where flexible functional methods are preferred. However, most functional clustering methods for longitudinal data do not allow controlling for the possible functional effects of covariates. Although traditional mixture-of-experts methods can include covariates and be extended to the functional setting, using nonlinear basis functions, satisfactory parsimonious functional methods required for robust functional coefficient estimation and clustering are still lacking. In this paper we propose a novel latent class functional mixed-effects model in which we assume the covariates have fixed functional effects, and the random curves follow a mixture of Gaussian processes that facilitates a model-based conditional clustering. A transformed penalized B-spline approach is employed for parsimonious modeling and robust model estimation. We propose a new iterative-REML method to choose the penalty parameters in heterogeneous data. The new method is applied to the latest data from the Religious Orders Study and Rush Memory and Aging Project, and four novel subtypes of cognitive changes are identified.

REFERENCES

ABRAHAM, C., CORNILLON, P. A., MATZNER-LØBER, E. and MOLINARI, N. (2003). Unsupervised curve clustering using B-splines. *Scand. J. Stat.* **30** 581–595. MR2002229 https://doi.org/10.1111/1467-9469.00350

BAR-JOSEPH, Z., GERBER, G., GIFFORD, D. K., JAAKKOLA, T. S. and SIMON, I. (2002). A new approach to analyzing gene expression time series data. In *Proceedings of the Sixth Annual International Conference on Computational Biology*. *RECOMB* '02 39–48. Association for Computing Machinery, New York, NY, USA.

BENNETT, D. A., BUCHMAN, A. S., BOYLE, P. A., BARNES, L. L., WILSON, R. S. and SCHNEIDER, J. A. (2018). Religious orders study and rush memory and aging project. *J. Alzheimer's Dis.* **64** 161–189.

BOYLE, P. A., WILSON, R. S., YU, L., BARR, A. M., HONER, W. G., SCHNEIDER, J. A. and BENNETT, D. A. (2013). Much of late life cognitive decline is not due to common neurodegenerative pathologies. *Ann. Neurol.* **74** 478–489.

BOYLE, P. A., YANG, J., YU, L., LEURGANS, S. E., CAPUANO, A. W., SCHNEIDER, J. A., WILSON, R. S. and BENNETT, D. A. (2017). Varied effects of age-related neuropathologies on the trajectory of late life cognitive decline. *Brain* **140** 804–812.

BOYLE, P. A., WANG, T., YU, L., WILSON, R. S., DAWE, R., ARFANAKIS, K., SCHNEIDER, J. A. and BENNETT, D. A. (2021). To what degree is late life cognitive decline driven by age-related neuropathologies? *Brain* **144** 2166–2175.

CHAMROUKHI, F. and NGUYEN, H. D. (2019). Model-based clustering and classification of functional data. *WIREs Data Mining and Knowledge Discovery* **9** e1298.

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

CHIOU, J.-M. and LI, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *J. Roy. Statist. Soc. Ser. B* **69** 679–699. MR2370075 https://doi.org/10.1111/j.1467-9868.2007.00605.x

Key words and phrases. Cohort study, functional mixed-effects model, latent classes, mixture Gaussian processes, penalized B-splines, ROSMAP studies.

COFFEY, N., HINDE, J. and HOLIAN, E. (2014). Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data. *Comput. Statist. Data Anal.* **71** 14–29. MR3131951 https://doi.org/10.1016/j.csda.2013.04.001

DELAIGLE, A., HALL, P. and PHAM, T. (2019). Clustering functional data into groups by using projections. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 271–304. MR3928143 https://doi.org/10.1111/rssb.12310

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

DODGE, H. H., WANG, C.-N., CHANG, C.-C. H. and GANGULI, M. (2011). Terminal decline and practice effects in older adults without dementia. *Neurology* **77** 722–730.

DU, P. and WANG, X. (2014). Penalized likelihood functional regression. *Statist. Sinica* **24** 1017–1041. MR3236451

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

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635 https://doi.org/10.1198/016214502760047131

GARCZAREK, U. M. (2002). Classification rules in standardized partition spaces. Dissertation, Univ. Dortmund. Available at http://hdl.handle.net/2003/2789.

GENOLINI, C., ECOCHARD, R., BENGHEZAL, M., DRISS, T., ANDRIEU, S. and SUBTIL, F. (2016). kmlShape: An efficient method to cluster longitudinal data (time-series) according to their shapes. *PLoS ONE* **11** 1–24.

GREEN, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **52** 443–452. MR1086796

GRÜN, B., SCHARL, T. and LEISCH, F. (2011). Modelling time course gene expression data with finite mixtures of linear additive models. *Bioinformatics* **28** 222–228.

GU, C. (1992). Cross-validating non-Gaussian data. *J. Comput. Graph. Statist.* **1** 169–179.

GU, C. and MA, P. (2005). Optimal smoothing in nonparametric mixed-effect models. *Ann. Statist.* **33** 1357–1379. MR2195638 https://doi.org/10.1214/009053605000000110

GUO, W. (2002). Functional mixed effects models. *Biometrics* **58** 121–128. MR1891050 https://doi.org/10.1111/j.0006-341X.2002.00121.x

HALL, C. B., LIPTON, R. B., SLIWINSKI, M. and STEWART, W. F. (2000). A change point model for estimating the onset of cognitive decline in preclinical Alzheimer's disease. *Stat. Med.* **19** 1555–1566.

HEARD, N. A., HOLMES, C. C. and STEPHENS, D. A. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *J. Amer. Statist. Assoc.* **101** 18–29. MR2252430 https://doi.org/10.1198/016214505000000187

HENDERSON, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31** 423–447.

JACK, C., KNOPMAN, D., JAGUST, W., PETERSEN, R., WEINER, M., AISEN, P., SHAW, L., VEMURI, P., WISTE, H. et al. (2013). Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. *Lancet Neurol.* **12** 207–216.

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). Model-based clustering for multivariate functional data. *Comput. Statist. Data Anal.* **71** 92–106. MR3131956 https://doi.org/10.1016/j.csda.2012.12.004

JAMES, G. M. and SILVERMAN, B. W. (2005). Functional adaptive model estimation. *J. Amer. Statist. Assoc.* **100** 565–576. MR2160560 https://doi.org/10.1198/016214504000001556

JAMES, G. M. and SUGAR, C. A. (2003). Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.* **98** 397–408. MR1995716 https://doi.org/10.1198/016214503000189

KONISHI, S., ANDO, T. and IMOTO, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* **91** 27–43. MR2050458 https://doi.org/10.1093/biomet/91.1.27

LUAN, Y. and LI, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics* **19** 474–482.

MA, P. and ZHONG, W. (2008). Penalized clustering of large-scale functional data with multiple covariates. *J. Amer. Statist. Assoc.* **103** 625–636. MR2435467 https://doi.org/10.1198/016214508000000247

MA, P., CASTILLO-DAVIS, C. I., ZHONG, W. and LIU, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Res.* **34** 1261–1269.

MARKESBERY, W. R. (2010). Neuropathologic alterations in mild cognitive impairment: A review. *J. Alzheimer's Dis.* **19** 221–228.

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models. Wiley Series in Probability and Statistics: Applied Probability and Statistics*. Wiley Interscience, New York. MR1789474 https://doi.org/10.1002/0471721182

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

MURPHY, K. and MURPHY, T. B. (2020). Gaussian parsimonious clustering models with covariates and a noise component. *Adv. Data Anal. Classif.* **14** 293–325. MR4118952 https://doi.org/10.1007/s11634-019-00373-8

PENG, J. and MÜLLER, H.-G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Ann. Appl. Stat.* **2** 1056–1077. MR2516804 https://doi.org/10.1214/08-AOAS172

PINHEIRO, J. C. and BATES, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

PROUST, C., JACQMIN-GADDA, H., TAYLOR, J. M. G., GANIAYRE, J. and COMMENGES, D. (2006). A non-linear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics* **62** 1014–1024. MR2297672 https://doi.org/10.1111/j.1541-0420.2006.00573.x

PROUST-LIMA, C., DARTIGUES, J.-F. and JACQMIN-GADDA, H. (2011). Misuse of the linear mixed model when evaluating risk factors of cognitive decline. *Am. J. Epidemiol.* **174** 1077–1088.

PROUST-LIMA, C., PHILIPPS, V. and LIQUET, B. (2017). Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *J. Stat. Softw.* **78** 1–56.

QIN, L.-X. and SELF, S. G. (2006). The clustering of regression models method with applications in gene expression data. *Biometrics* **62** 526–533. MR2236835 https://doi.org/10.1111/j.1541-0420.2005.00498.x

RAMONI, M. F., SEBASTIANI, P. and KOHANE, I. S. (2002). Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. USA* **99** 9121–9126. MR1909705 https://doi.org/10.1073/pnas.132656399

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

REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. MR0738930 https://doi.org/10.1137/1026034

RICE, J. A. and WU, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57** 253–259. MR1833314 https://doi.org/10.1111/j.0006-341X.2001.00253.x

RODRIGUEZ, A. and DUNSON, D. B. (2014). Functional clustering in nested designs: Modeling variability in reproductive epidemiology studies. *Ann. Appl. Stat.* **8** 1416–1442. MR3271338 https://doi.org/10.1214/14-AOAS751

ROTHENBERG, T. J. (1971). Identification in parametric models. *Econometrica* **39** 577–591. MR0436944 https://doi.org/10.2307/1913267

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

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

SERBAN, N. and WASSERMAN, L. (2005). CATS: Clustering after transformation and smoothing. *J. Amer. Statist. Assoc.* **100** 990–999. MR2201025 https://doi.org/10.1198/016214504000001574

SHI, M., WEISS, R. E. and TAYLOR, J. M. G. (1996). An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Appl. Stat.* **45** 151–163.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J. Roy. Statist. Soc. Ser. B* **47** 1–52. MR0805063

STEINERMAN, J. R., HALL, C. B., SLIWINSKI, M. J. and LIPTON, R. B. (2010). Modeling cognitive trajectories within longitudinal studies: A focus on older adults. *J. Amer. Geriatr. Soc.* **58** S313–S318.

STERN, Y., BARNES, C. A., GRADY, C., JONES, R. N. and RAZ, N. (2019). Brain reserve, cognitive reserve, compensation, and maintenance: Operationalization, validity, and mechanisms of cognitive resilience. *Neurobiol. Aging* **83** 124–129.

TARPEY, T. (2007). Linear transformations and the k-means clustering algorithm: Applications to clustering curves. *Amer. Statist.* **61** 34–40. MR2339145 https://doi.org/10.1198/000313007X171016

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, Chichester. MR0838090

VERBEKE, G. and LESAFFRE, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *J. Amer. Statist. Assoc.* **91** 217–221.

WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40** 364–372. MR0522220

WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150. MR0701084

WAHBA, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13** 1378–1402. MR0811498 https://doi.org/10.1214/aos/1176349743

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

WAKEFIELD, J. C., ZHOU, C. and SELF, S. G. (2003). Modelling gene expression data over time: Curve clustering with informative prior distributions. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) (J. Bernardo, M. Bayarri,

J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West, eds.) 721–732. Oxford Univ. Press, New York. MR2003536

WANG, Y. (1998). Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.* **93** 341–348.

WANG, T., LEI, Y., LEURGANS, S. E., WILSON, R. S., BENNETT, D. A. and BOYLE, P. A. (2022). Supplement to "Conditional functional clustering for longitudinal data with heterogeneous nonlinear patterns." https://doi.org/10.1214/21-AOAS1542SUPP

WILSON, R., BECKETT, L., BARNES, L., SCHNEIDER, J., BACH, J., EVANS, D. and BENNETT, D. (2002). Individual differences in rates of change in cognitive abilities of older persons. *Psychology and Aging* **17** 179–193.

WOOD, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Assoc.* **99** 673–686. MR2090902 https://doi.org/10.1198/016214504000000980

YAO, F., FU, Y. and LEE, T. C. M. (2010). Functional mixture regression. *Biostatistics* **12** 341–353.

ZHU, X. and QU, A. (2018). Cluster analysis of longitudinal profiles with subgroups. *Electron. J. Stat.* **12** 171–193. MR3756096 https://doi.org/10.1214/17-EJS1389

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1215–1235
https://doi.org/10.1214/21-AOAS1543

IMPACT EVALUATION OF THE LAPD COMMUNITY SAFETY PARTNERSHIP

BY SYDNEY KAHMANN[1,a], ERIN HARTMAN[2,b], JORJA LEAP[3,c] AND P. JEFFREY BRANTINGHAM[4,d]

[1]*Department of Statistics, University of California, Los Angeles,* [a]*skahmann@ucla.edu*
[2]*Department of Political Science, University of California, Berkeley ,* [b]*ekhartman@berkeley.edu*
[3]*Department of Social Welfare, University of California, Los Angeles,* [c]*jleap@ucla.edu*
[4]*Department of Anthropology, University of California, Los Angeles,* [d]*branting@ucla.edu*

In 2011, the Los Angeles Police Department (LAPD), in conjunction with other governmental and nonprofit groups, launched the Community Safety Partnership (CSP) in several public housing developments in Los Angeles. Following a relationship-based policing model, officers were assigned to work collaboratively with community members to reduce crime and build trust. However, evaluating the causal impact of this policy intervention is difficult, given the notable differences between communities where CSP was implemented and the surrounding communities in South Los Angeles. In this paper we use a novel data set, based on the LAPD's reported crime incidents and calls-for-service, to evaluate the effectiveness of this program via augmented synthetic control models, a cutting-edge method for policy evaluation. We perform falsification analyses to evaluate the robustness of the results. In the public housing developments where it was first deployed, we find that CSP exhibited modest but statistically insignificant reductions in reported violent crime incidents, shots fired and violent crime calls-for-service, and Part I reported crime incidents. We do not find evidence of crime displacement from CSP regions to neighboring control regions.

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., DIAMOND, A. and HAINMUELLER, J. (2015). Comparative politics and the synthetic control method. *Amer. J. Polit. Sci.* **59** 495–510.

ABADIE, A. and GARDEAZABAL, J. (2003). The economic costs of conflict: A case study of the Basque country. *Am. Econ. Rev.* **93** 113–132.

ABADIE, A. and L'HOUR, J. (2020). A penalized synthetic control estimator for disaggregated data.

ANGRIST, J. D. and PISCHKE, J.-S. (2008). *Mostly Harmless Econometrics*: *An Empiricist's Companion*. Princeton University Press, Princeton.

ARKHANGELSKY, D., ATHEY, S., HIRSHBERG, D. A., IMBENS, G. W. and WAGER, S. (2018). Synthetic difference in differences. ArXiv.

ATHEY, S. and IMBENS, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *J. Econ. Perspect.* **31** 3–32.

ATHEY, S. and IMBENS, G. (2018). Design-based analysis in difference-in-differences settings with staggered adoption. ArXiv.

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

BEN-MICHAEL, E., FELLER, A. and ROTHSTEIN, J. (2019). Synthetic controls with staggered adoption. ArXiv.

BEN-MICHAEL, E., FELLER, A. and ROTHSTEIN, J. (2021). The augmented synthetic control method. *J. Amer. Statist. Assoc.* 1–34.

BLACKSTONE, J. (2014). A change of tune for LAPD as community policing makes gains.

Key words and phrases. Gang violence, community policing, time-series cross-sectional data, augmented synthetic control method, policy evaluation, causal inference.

BLAIR, G., WEINSTEIN, J., CHRISTIA, F., ARIAS, E., BADRAN, E., BLAIR, R. A., CHEEMA, A., FAROOQUI, A., FETZER, T. et al. (2021). Does community policing build trust in police and reduce crime? Evidence from six coordinated field experiments in the global South. Submitted.

BLUMSTEIN, A. (2018). Science and technology and the president's crime commission: Past and future. *Criminol. Public Policy* **17** 271–282.

BLUMSTEIN, A., WALLMAN, J. and FARRINGTON, D. (2006). *The Crime Drop in America*. Cambridge Univ. Press, Cambridge.

BOTOSARU, I. and FERMAN, B. (2019). On the role of covariates in the synthetic control method. *Econom. J.* **22** 117–130. MR4021116 https://doi.org/10.1093/ectj/utz001

BRAGA, A. A. and WEISBURD, D. L. (2012). The effects of focused deterrence strategies on crime: A systematic review and meta-analysis of the empirical evidence. *J. Res. Crime Delinq.* **49** 323–358.

BRAGA, A. A., KENNEDY, D. M., WARING, E. J. and PIEHL, A. M. (2001). Problem-oriented policing, deterrence, and youth violence: An evaluation of Boston's operation ceasefire. *J. Res. Crime Delinq.* **38** 195–225.

BRANTINGHAM, P. J., TITA, G. and HERZ, D. (2021). The impact of the city of Los Angeles mayor's office of Gang Reduction and Youth Development (GRYD) comprehensive strategy on crime in the city of Los Angeles. *Justice Eval. J.* 1–20.

CALLAWAY, B. and SANT'ANNA, P. H. C. (2020). Difference-in-differences with multiple time periods. *J. Econometrics*.

CARD, D. and KRUEGER, A. B. (1993). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania Technical Report National Bureau of Economic Research.

CHERNOZHUKOV, V., WÜTHRICH, K. and ZHU, Y. (2021). An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls.

CORDNER, G. W. (1997). Community policing: Elements and effects. *Crit. Issues Policing: Contemp. Read.* **5** 401–418.

DOJ, U. (2009). Community policing defined. *Office Community Oriented Policing Serv*.

DOUDCHENKO, N. and IMBENS, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. ArXiv.

ECK, J. E. and MAGUIRE, E. R. (2000). Have changes in policing reduced violent crime? An assessment of the evidence. *Crime Drop Amer.* **207** 207–265.

EGAMI, N. and YAMAUCHI, S. (2021). Using multiple pre-treatment periods to improve difference-in-differences and staggered adoption design. ArXiv.

FAGAN, J. and MACDONALD, J. (2012). Policing, crime and legitimacy in New York and Los Angeles: The social and political contexts of two historic crime declines. Columbia Public Law Research Paper 12-315.

FERMAN, B. and PINTO, C. (2019). Synthetic controls with imperfect pre-treatment fit. ArXiv.

GOLDSTEIN, H. (1977). *Policing a Free Society*. Ballinger Pub. Co., Cambridge, MA.

GOLDSTEIN, H. (1979). Improving policing: A problem-oriented approach. *Crime Delinq.* **25** 236–258.

GROGGER, J. (2002). The effects of civil gang injunctions on reported violent crime: Evidence from Los Angeles county. *J. Law Econ.* **45** 69–90.

HARTMAN, E. and HIDALGO, F. D. (2018). An equivalence approach to balance and placebo tests. *Amer. J. Polit. Sci.* **62** 1000–1013.

HAZLETT, C. and XU, Y. (2018). Trajectory balancing: A general reweighting approach to causal inference with time-series cross-sectional data.

HECKMAN, J. J. and HOTZ, V. J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *J. Amer. Statist. Assoc.* **84** 862–874.

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

IMAI, K., KIM, I. and WANG, E. (2020). Matching methods for causal inference with time-series cross-sectional data. Unpublished Manuscript.

KAHMANN, S., HARTMAN, E., LEAP, J. and BRANTINGHAM, P. J (2022). Supplement to "Impact Evaluation of the LAPD Community Safety Partnership." https://doi.org/10.1214/21-AOAS1543SUPP

KATZENBACH, N. D. (1967). *The Challenge of Crime in a Free Society: A Report by the President's Commission on Law Enforcement and Administration of Justice*. United States Government Printing Office.

KELLING, G. L. and MOORE, M. H. (1989). *The Evolving Strategy of Policing* **4**. US Department of Justice, Office of Justice Programs, National Institute of Justice.

KELLING, G. L., PATE, A., FERRARA, A., UTNE, M. and BROWN, C. E. (1981). *The Newark Foot Patrol Experiment*. Police Foundation, Washington, DC, 94–96.

KENNEDY, D. M. (1997). Pulling levers: Chronic offenders, high-crime settings, and a theory of prevention. *Valparaiso Univ. Law Rev.* **31** 449–484.

KENNEDY, D., PIEHL, A. and BRAGA, A. (1996). Youth violence in Boston: Gun markets, serious youth offenders, and a use reduction strategy. *Law Contemp. Probl.* **59** 147–183.

KLINGER, D. A. and BRIDGES, G. S. (1997). Measurement error in calls-for-service as an indicator of crime. *Criminology* **35** 705–726.

LEAP, J. (2020). *Evaluation of the LAPD Community Safety Partnership*. UCLA Luskin School of Public Affairs, Los Angeles.

MACDONALD, J. M. (2002). The effectiveness of community policing in reducing urban violence. *Crime Delinq.* **48** 592–618.

NAHUM-SHANI, I., SMITH, S. N., SPRING, B. J., COLLINS, L. M., WITKIEWITZ, K., TEWARI, A. and MURPHY, S. A. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals Behav. Med.* **52** 446–462.

O'NEILL, S., KREIF, N., GRIEVE, R., SUTTON, M. and SEKHON, J. S. (2016). Estimating causal effects: Considering three alternatives to difference-in-differences estimation. *Health Serv. Outcomes Res. Methodol.* **16** 1–21. https://doi.org/10.1007/s10742-016-0146-8

FEDERAL BUREAU OF INVESTIGATION (2021). Uniform Crime Reporting Program Data: Supplementary Homicide Reports, 1987–2017. Inter-university Consortium for Political and Social Research.

LA OFFICE OF THE MAYOR (2017). Mayor Garcetti announces new expansion of Community Safety Partnership.

PARK, J., SCHOENBERG, F. P., BERTOZZI, A. L. and BRANTINGHAM, P. J. (2021). Investigating clustering and violence interruption in gang-related violent crime data using spatial–temporal point processes with covariates. *J. Amer. Statist. Assoc.* 1–14.

PORTER, L. C., CURTIS, A., JEFFERIS, E. and MITCHELL, S. (2019). Where's the crime? Exploring divergences between call data and perceptions of local crime. *Br. J. Criminol.* **60** 444–467.

REISIG, M. D. (2010). Community and problem-oriented policing. *Crime Justice* **39** 1–53.

LAPD NEWS RELEASE (2015). LAPD's Community Safety Partnership program NR15021SF.

RICE, C. and LEE, S. K. (2015). *Relationship-Based Policing Achieving Safety in Watts*. Advancement Project, Washington, DC.

RIDGEWAY, G., GROGGER, J., MOYER, R. A. and MACDONALD, J. M. (2019). Effect of gang injunctions on crime: A study of Los Angeles from 1988–2014. *J. Quant. Criminol.* **35** 517–541.

ROBBINS, M. W., SAUNDERS, J. and KILMER, B. (2017). A framework for synthetic control methods with high-dimensional, micro-level data: Evaluating a neighborhood-specific crime intervention. *J. Amer. Statist. Assoc.* **112** 109–126. MR3646556 https://doi.org/10.1080/01621459.2016.1213634

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

SHERMAN, L. W., MILTON, C., KELLY, T. V. and MACBRIDE, T. F. (1973). *Team Policing: Seven Case Studies*. Police Foundation, Washington, DC.

SIEGLER, K. (2013). After years of violence, L.A.'s Watts sees crime subside.

SKOGAN, W. G. (2006). *Police and Community in Chicago: A Tale of Three Cities*. Oxford Univ. Press, Oxford.

STREETER, K. (2014). In Jordan Downs housing project, police are forging a new relationship.

TREMBLAY, A., HERZ, D., ZACHERY, R. and KRAUS, M. (2020). The Los Angeles Mayor's Office of Gang Reduction and Youth Development Comprehensive Strategy. GRYD Research Brief No. 1.

The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1236–1252
https://doi.org/10.1214/21-AOAS1544

HIGHER CRITICISM FOR DISCRIMINATING WORD-FREQUENCY TABLES AND AUTHORSHIP ATTRIBUTION

BY ALON KIPNIS[a]

Department of Statistics, Stanford University, [a]*kipnisal@stanford.edu*

We adapt the higher criticism (HC) goodness-of-fit test to measure the closeness between word-frequency tables. We apply this measure to authorship attribution challenges, where the goal is to identify the author of a document using other documents whose authorship is known. The method is simple yet performs well without handcrafting and tuning, reporting accuracy at the state-of-the-art level in various current challenges. As an inherent side effect, the HC calculation identifies a subset of discriminating words. In practice, the identified words have low variance across documents belonging to a corpus of homogeneous authorship. We conclude that in comparing the similarity of a new document and a corpus of a single author, HC is mostly affected by words characteristic of the author and is relatively unaffected by topic structure.

REFERENCES

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

ARIAS-CASTRO, E. and WANG, M. (2015). The sparse Poisson means model. *Electron. J. Stat.* **9** 2170–2201. MR3406276 https://doi.org/10.1214/15-EJS1066

BALAKRISHNAN, S. and WASSERMAN, L. (2018). Hypothesis testing for high-dimensional multinomials: A selective review. *Ann. Appl. Stat.* **12** 727–749. MR3834283 https://doi.org/10.1214/18-AOAS1155SF

BALAKRISHNAN, S. and WASSERMAN, L. (2019). Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *Ann. Statist.* **47** 1893–1927. MR3953439 https://doi.org/10.1214/18-AOS1729

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis*: *Theory and Practice*. The MIT Press, Cambridge, MA–London. With the collaboration of Richard J. Light and Frederick Mosteller. MR0381130

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

BRESLOW, N. E. (1984). Extra-Poisson variation in log-linear models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **33** 38–44.

BROWN, L. D., ZHANG, R. and ZHAO, L. (2001). Root un-root methodology for nonparametric density estimation. Techical Report, The Wharton School, Univ. Pennsylvania.

CAI, T. T., JENG, X. J. and JIN, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 629–662. MR2867452 https://doi.org/10.1111/j.1467-9868.2011.00778.x

CAI, T. T., JIN, J. and LOW, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35** 2421–2449. MR2382653 https://doi.org/10.1214/009053607000000334

CHANG, J. and BLEI, D. M. (2010). Hierarchical relational models for document networks. *Ann. Appl. Stat.* **4** 124–150. MR2758167 https://doi.org/10.1214/09-AOAS309

CHURCH, K. W. and GALE, W. A. (1995). Poisson mixtures. *Nat. Lang. Eng.* **1** 163–190.

COX, D. R. and BRANDWOOD, L. (1959). On a discriminatory problem connected with the works of Plato. *J. Roy. Statist. Soc. Ser. B* **21** 195–200. MR0109102

CRESSIE, N. and READ, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46** 440–464. MR0790631

DELAIGLE, A. and HALL, P. (2009). Higher criticism in the context of unknown distribution, non-independence and classification. In *Perspectives in Mathematical Sciences. I. Stat. Sci. Interdiscip. Res.* **7** 109–138. World Sci. Publ., Hackensack, NJ. MR2581742 https://doi.org/10.1142/9789814273633_0006

Key words and phrases. Higher criticism, two-sample testing, nonparametric methods, unsupervised learning, feature selection, authorship attribution.

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
DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265
DONOHO, D. and JIN, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **105** 14790–14795.
DONOHO, D. and JIN, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4449–4470. With electronic supplementary materials available online. MR2546396 https://doi.org/10.1098/rsta.2009.0129
DONOHO, D. and JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.* **30** 1–25. MR3317751 https://doi.org/10.1214/14-STS506
DONOHO, D. L. and KIPNIS, A. (2020). Higher criticism to compare two large frequency tables, with sensitivity to possible rare and weak differences.
EFRON, B. and THISTED, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63** 435–447.
GLICKMAN, M., BROWN, J. and SONG, R. (2019). (A) data in the life: Authorship attribution in Lennon–McCartney songs. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.130f856e
GRIFFITHS, T. L., JORDAN, M. I., TENENBAUM, J. B. and BLEI, D. M. (2004). Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems* 17–24.
HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357 https://doi.org/10.1214/09-AOS764
HAMILTON, A., MADISON, J. and JAY, J. (1961). The federalist papers, ed. Clinton Rossiter (New York: New American Library, 1961), 301. *Federalism*, *Citizenship*, *and Community* **207**.
HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Stat.* **36** 369–408. MR0173322 https://doi.org/10.1214/aoms/1177700150
HOLMES, D. I. (1985). The analysis of literary style—a review. *J. R. Stat. Soc., A* **148** 328–341.
INGSTER, Y. I., TSYBAKOV, A. B. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526. MR2747131 https://doi.org/10.1214/10-EJS589
JAGER, L. and WELLNER, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Ann. Statist.* **35** 2018–2053. MR2363962 https://doi.org/10.1214/0009053607000000244
JIN, J. and KE, Z. T. (2016). Rare and weak effects in large-scale inference: Methods and phase diagrams. *Statist. Sinica* **26** 1–34. MR3468343
JIN, J. and WANG, W. (2016). Influential features PCA for high dimensional clustering. *Ann. Statist.* **44** 2323–2359. MR3576543 https://doi.org/10.1214/15-AOS1423
JUOLA, P. (2008). Authorship attribution. *Found. Trends Inf. Retr.* **1** 233–334.
KESTEMONT, M., TSCHUGGNALL, M., STAMATATOS, E., DAELEMANS, W., SPECHT, G., STEIN, B. and POTTHAST, M. (2018). Overview of the author identification task at PAN-2018: Cross-domain authorship attribution and style change detection. In *Working Notes Papers of the CLEF* 2018 *Evaluation Labs. Avignon*, *France*, *September* 10–14, 2018/*Cappellato*, *Linda* [*edit.*]; *et al.* 1–25.
KESTEMONT, M., MANJAVACAS, E., MARKOV, I., BEVENDORFF, J., WIEGMANN, M., STAMATATOS, E., POTTHAST, M. and STEIN, B. (2020). Overview of the cross-domain authorship verification task at PAN 2020. In *CLEF* (*Working Notes*).
KIPNIS, A. (2020). Higher criticism as an unsupervised authorship discriminator. In *CLEF* (*Working Notes*).
KIPNIS, A. (2022). Supplement to "Higher criticism for discriminating word-frequency tables and authorship attribution." https://doi.org/10.1214/21-AOAS1544SUPP
LEHMANN, E. L. (1975). *Nonparametrics*: *Statistical Methods Based on Ranks*. *Holden-Day Series in Probability and Statistics*. Holden-Day, Inc., San Francisco, CA; McGraw-Hill International Book Co., New York–Düsseldorf. With the special assistance of H. J. M. d'Abrera. MR0395032
LI, J. and SIEGMUND, D. (2015). Higher criticism: *p*-values and criticism. *Ann. Statist.* **43** 1323–1350. MR3346705 https://doi.org/10.1214/15-AOS1312
MANNING, C., RAGHAVAN, P. and SCHÜTZE, H. (2010). Introduction to information retrieval. *Nat. Lang. Eng.* **16** 100–103.
MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. *Monographs on Statistics and Applied Probability*. CRC Press, London. Second edition [of MR0727836]. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6
MOSTELLER, F. and WALLACE, D. L. (1963). Inference in an authorship problem. *J. Amer. Statist. Assoc.* **58** 275–309.
MOSTELLER, F. and WALLACE, D. L. (1984). *Applied Bayesian and Classical Inference*: *The Case of The Federalist Papers*. *Springer Series in Statistics*. Springer, New York. Second edition of *Inference and disputed authorship: the Federalist*. MR0766742 https://doi.org/10.1007/978-1-4612-5256-6

MUKHERJEE, R., PILLAI, N. S. and LIN, X. (2015). Hypothesis testing for high-dimensional sparse binary regression. *Ann. Statist.* **43** 352–381. MR3311863 https://doi.org/10.1214/14-AOS1279

NORVIG, P. (2013). Common words in Google books. http://norvig.com/mayzner.html.

PEARSON, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag.* **50** 157–175.

QI, P., DOZAT, T., ZHANG, Y. and MANNING, C. D. (2018). Universal dependency parsing from scratch. In *Proceedings of the CoNLL* 2018 *Shared Task*: *Multilingual Parsing from Raw Text to Universal Dependencies* 160–170. Association for Computational Linguistics, Brussels, Belgium.

READ, T. R. and CRESSIE, N. A. (2012). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer Science & Business Media. MR0955054 https://doi.org/10.1007/978-1-4612-4578-0

ROBERTS, M. E., STEWART, B. M. and AIROLDI, E. M. (2016). A model of text for experimentation in the social sciences. *J. Amer. Statist. Assoc.* **111** 988–1003. MR3561924 https://doi.org/10.1080/01621459.2016.1141684

ROSS, G. J. (2020). Tracking the evolution of literary style via Dirichlet-multinomial change point regression. *J. Roy. Statist. Soc. Ser. A* **183** 149–167. MR4049658

SICHEL, H. S. (1974). On a distribution representing sentence-length in written prose. *J. R. Stat. Soc., A* **137** 25–34.

THISTED, R. and EFRON, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika* **74** 445–455. MR0909350 https://doi.org/10.1093/biomet/74.3.445

TILAHUN, G., FEUERVERGER, A. and GERVERS, M. (2012). Dating medieval English charters. *Ann. Appl. Stat.* **6** 1615–1640. MR3058677 https://doi.org/10.1214/12-AOAS566

WAKE, W. C. (1957). Sentence-length distributions of Greek authors. *J. R. Stat. Soc., A* **120** 331–346.

ZHENG, R., LI, J., CHEN, H. and HUANG, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *J. Am. Soc. Inf. Sci. Technol.* **57** 378–393.