

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

- A Bayesian precision medicine framework for calibrating individualized therapeutic indices in cancer
ABHISEK SAHA, MIN JIN HA,
SATWIK ACHARYYA AND VEERABHADRAN BALADANDAYUTHAPANI 2055
- Semiparametric Bayesian forecasting of SpatioTemporal earthquake occurrences
GORDON J. ROSS AND ALEKSANDAR A. KOLEV 2083
- Estimating the stillbirth rate for 195 countries using a Bayesian sparse regression model with
temporal smoothing ZHENGFAN WANG, MIRANDA J. FIX, LUCIA HUG, ANU MISHRA,
DANZHEN YOU, HANNAH BLENCOWE, JON WAKEFIELD AND LEONTINE ALKEMA 2101
- Functional random effects modeling of brain shape and connectivity
EARDI LILA AND JOHN A. D. ASTON 2122
- Hierarchical resampling for bagging in multistudy prediction with applications to human
neurochemical sensing GABRIEL LOEWINGER, PRASAD PATIL,
KENNETH T. KISHIDA AND GIOVANNI PARMIGIANI 2145
- Network differential connectivity analysis SEN ZHAO AND ALI SHOJAIE 2166
- Estimating heterogeneous gene regulatory networks from zero-inflated single-cell expression data
QIUYU WU AND XIANGYU LUO 2183
- Accounting for survey design in Bayesian disaggregation of survey-based areal estimates of
proportions: An application to the American Community Survey MARCO H. BENEDETTI,
VERONICA J. BERROCAL AND RODERICK J. LITTLE 2201
- Bayesian multivariate sparse functional principal components analysis with application to
longitudinal microbiome multiomics data LINGJING JIANG, CHRIS ELROD, JANE J. KIM,
AUSTIN D. SWAFFORD, ROB KNIGHT AND WESLEY K. THOMPSON 2231
- Data-adaptive efficient estimation strategies for biomarker studies embedded in randomized trials
WEI ZHANG, ZHIWEI ZHANG, JAMES F. TROENDLE AND AIYI LIU 2250
- An omnibus test for detection of subgroup treatment effects via data partitioning
YIFEI SUN, XUMING HE AND JIANHUA HU 2266
- Assessing treatment effect through compliance score in randomized trials with noncompliance
ZONGHUI HU, ZHIWEI ZHANG AND DEAN FOLLMANN 2279
- Spatiotemporal satellite data imputation using sparse functional data analysis
WEICHENG ZHU, ZHENGYUAN ZHU AND XIONGTAO DAI 2291
- Multilevel time-series models for small area estimation at different frequencies and domain levels
HARM JAN BOONSTRA AND JAN VAN DEN BRAKEL 2314
- How many refugees and migrants died trying to reach Europe? Joint population size and total
estimation ALESSIO FARCOMENI 2339
- Full Bayesian inference in hidden Markov models of plant growth
GAUTIER VIAUD, YUTING CHEN AND PAUL-HENRY COURNÈDE 2352
- Extended stochastic block models with application to criminal networks
SIRIO LEGRAMANTI, TOMMASO RIGON, DANIELE DURANTE AND DAVID B. DUNSON 2369

continued

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—Continued from front cover

Two-sample tests for multivariate repeated measurements of histogram objects with applications to wearable device data	JINGRU ZHANG, KATHLEEN R. MERIKANGAS, HONGZHE LI AND HAOCHANG SHOU	2396
Parsimonious Bayesian factor analysis for modelling latent structures in spectroscopy data	ALESSANDRO CASA, TOM F. O'CALLAGHAN AND THOMAS BRENDAN MURPHY	2417
Phylogenetically informed Bayesian truncated copula graphical models for microbial association networks	HEE CHEOL CHUNG, IRINA GAYNANOVA AND YANG NI	2437
Causal inference for the effect of mobility on COVID-19 deaths	MATTEO BONVINI, EDWARD H. KENNEDY, VALERIE VENTURA AND LARRY WASSERMAN	2458
Bayesian hierarchical random-effects meta-analysis and design of phase I clinical trials	RUITAO LIN, HAOLUN SHI, GUOSHENG YIN, PETER F. THALL, YING YUAN AND CHRISTOPHER R. FLOWERS	2481
Estimating functional parameters for understanding the impact of weather and government interventions on COVID-19 outbreak	CHIH-LI SUNG	2505
Clustering and forecasting multiple functional time series	CHEN TANG, HAN LIN SHANG AND YANRONG YANG	2523
Semi-supervised nonparametric Bayesian modelling of spatial proteomics	OLIVER M. CROOK, KATHRYN S. LILLEY, LAURENT GATTO AND PAUL D. W. KIRK	2554
Bayesian data synthesis and the utility-risk trade-off for mixed epidemiological data	JOSEPH FELDMAN AND DANIEL R. KOWAL	2577
Hierarchical Bayesian modeling of ocean heat content and its uncertainty	SAMUEL BAUGH AND KAREN MCKINNON	2603
Bayesian inference for brain activity from functional magnetic resonance imaging collected at two spatial resolutions	ANDREW S. WHITEMAN, ANDREAS J. BARTSCH, JIAN KANG AND TIMOTHY D. JOHNSON	2626
Testing for differential abundance in compositional counts data, with application to microbiome studies	BARAK BRILL, AMNON AMIR AND RUTH HELLER	2648
Mapping interstellar dust with Gaussian processes	ANDREW C. MILLER, LAUREN ANDERSON, BORIS LEISTEDT, JOHN P. CUNNINGHAM, DAVID W. HOGG AND DAVID M. BLEI	2672
Modelling extremes of spatial aggregates of precipitation using conditional methods	JORDAN RICHARDS, JONATHAN A. TAWN AND SIMON BROWN	2693
A spatial causal analysis of wildland fire-contributed PM _{2.5} using numerical model output	ALEXANDRA LARSEN, SHU YANG, BRIAN J. REICH AND ANA G. RAPPOLD	2714
Dimensions, power and factors in an observational study of behavioral problems after physical abuse of children	TING YE, DYLAN S. SMALL AND PAUL R. ROSENBAUM	2732
Scalar on network regression via boosting	EMILY L. MORRIS, KEVIN HE AND JIAN KANG	2755
Multiscale spectral modelling for nonstationary time series within an ordered multiple-trial experiment	JONATHAN EMBLETON, MARINA I. KNIGHT AND HERNANDO OMBAO	2774
Bayesian bi-clustering methods with applications in computational biology	HAN YAN, JIEXING WU, YANG LI AND JUN S. LIU	2804

THE ANNALS OF APPLIED STATISTICS

Vol. 16, No. 4, pp. 2055–2831 December 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

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

President-Elect: Michael Kosorok, Department of Biostatistics and Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599, USA

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

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

Treasurer: Jiashun Jin, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA

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

IMS PUBLICATIONS

The Annals of Statistics. *Editors:* Enno Mammen, Institute for Applied Mathematics, Heidelberg University, 69120 Heidelberg, Germany. Lan Wang, Miami Herbert Business School, University of Miami, Coral Gables, FL 33124, USA

The Annals of Applied Statistics. *Editor-In-Chief:* Ji Zhu, Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

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

The Annals of Applied Probability. *Editors:* Kavita Ramanan, Division of Applied Mathematics, Brown University, Providence, RI 02912, USA. Qi-Man Shao, Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, Guangdong 518055, P.R. China

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

The IMS Bulletin. *Editor:* Tati Howell, bulletin@imstat.org

***The Annals of Applied Statistics* [ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 16, Number 4, December 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.

A BAYESIAN PRECISION MEDICINE FRAMEWORK FOR CALIBRATING INDIVIDUALIZED THERAPEUTIC INDICES IN CANCER

BY ABHISEK SAHA^{1,a}, MIN JIN HA^{2,b}, SATWIK ACHARYYA^{3,c} AND VEERABHADRAN BALADANDAYUTHAPANI^{3,d}

¹*Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institute of Health,*
^a*abhisek.saha@uconn.edu*

²*Department of Biostatistics, Graduate School of Public Health, Yonsei University, Seoul, Korea,* ^b*minjin.ha@gmail.com*

³*Department of Biostatistics, University of Michigan,* ^c*satwika@umich.edu*, ^d*veerab@umich.edu*

The development and clinical implementation of evidence-based precision medicine strategies has become a realistic possibility, primarily due to the rapid accumulation of large-scale genomics and pharmacological data from diverse model systems: patients, cell lines and drug perturbation studies. We introduce a novel Bayesian modeling framework called the individualized therapeutic index (iR_X) model to integrate high-throughput pharmacogenomic data across model systems. Our iR_X model achieves three main goals: first, it exploits the conserved biology between patients and cell lines to calibrate therapeutic response of drugs in patients; second, it finds optimal cell line avatars as proxies for patient(s); and finally, it identifies key genomic drivers explaining cell line-patient similarities. This is achieved through a semi-supervised learning approach that conflates (unsupervised) sparse latent factor models with (supervised) penalized regression techniques. We propose a unified and tractable Bayesian model for estimation, and inference is conducted via efficient posterior sampling schemes. We illustrate and validate our approach using two existing clinical trial data sets in multiple myeloma and breast cancer studies. We show that our iR_X model improves prediction accuracy compared to naive alternative approaches, and it consistently outperforms existing methods in literature in both multiple simulation scenarios as well as real clinical examples.

REFERENCES

- ARJMAND, B., GOODARZI, P., MOHAMADI-JAHANI, F., FALAHZADEH, K. and LARIJANI, B. (2017). Personalized regenerative medicine. *Acta Med. Iran.* **55** 144–149.
- ASSMANN, C., BOYSEN-HOGREFE, J. and PAPE, M. (2016). Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem. *J. Econometrics* **192** 190–206. [MR3463672 https://doi.org/10.1016/j.jeconom.2015.10.010](https://doi.org/10.1016/j.jeconom.2015.10.010)
- BARRETINA, J., CAPONIGRO, G., STRANSKY, N., VENKATESAN, K., MARGOLIN, A. A., KIM, S., WILSON, C. J., LEHÁR, J., KRYUKOV, G. V. et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483** 603.
- BEZDEK, J. C., EHRLICH, R. and FULL, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Comput. Geosci.* **10** 191–203.
- BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. [MR2806429 https://doi.org/10.1093/biomet/asr013](https://doi.org/10.1093/biomet/asr013)
- BHATTACHARYYA, R., HA, M. J., LIU, Q., AKBANI, R., LIANG, H. and BALADANDAYUTHAPANI, V. (2020). Personalized network modeling of the pan-cancer patient and cell line interactome. *JCO Clin. Cancer Inform.* **4** 399–411.
- BINNEMARS-POSTMA, K., BANSAL, R., STORM, G. and PRAKASH, J. (2017). Targeting the Stat6 pathway in tumor-associated macrophages reduces tumor growth and metastatic niche formation in breast cancer. *FASEB J.* **32** 969–978.

Key words and phrases. Bayesian methods, genomic data integration, high-dimensional regression, latent factor models, precision medicine, semi-supervised learning.

- BOONSTRA, P. S., MUKHERJEE, B. and TAYLOR, J. M. G. (2013). Bayesian shrinkage methods for partially observed data with many predictors. *Ann. Appl. Stat.* **7** 2272–2292. [MR3161722](#) <https://doi.org/10.1214/13-AOAS668>
- BØVELSTAD, H. M., NYGÅRD, S., STØRVOLD, H. L., ALDRIN, M., BORGAN, Ø., FRIGESSI, A. and LINGJÆRDE, O. C. (2007). Predicting survival from microarray data—a comparative study. *Bioinformatics* **23** 2080–2087.
- BRODACZEWSKA, K. K., SZCZYLIK, C., FIEDOROWICZ, M., PORTA, C. and CZARNECKA, A. M. (2016). Choosing the right cell line for renal cell cancer research. *Mol. Cancer* **15** 83. <https://doi.org/10.1186/s12943-016-0565-8>
- BROOKS, S. P. and GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7** 434–455. [MR1665662](#) <https://doi.org/10.2307/1390675>
- BROOKS, A. D., JACOBSEN, K. M., LI, W., SHANKER, A. and SAYERS, T. J. (2010). Bortezomib sensitizes human renal cell carcinomas to TRAIL apoptosis through increased activation of caspase-8 in the death-inducing signaling complex. *Mol. Cancer Res.* **8** 729–738.
- BYRNE, A. T., ALFÉREZ, D. G., AMANT, F., ANNIBALI, D., ARRIBAS, J., BIANKIN, A. V., BRUNA, A., BUDINSKÁ, E., CALDAS, C. et al. (2017). Interrogating open issues in cancer precision medicine with patient-derived xenografts. *Nat. Rev. Cancer*.
- CALIFANO, A. and ALVAREZ, M. J. (2017). The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nat. Rev. Cancer* **17** 116–130. <https://doi.org/10.1038/nrc.2016.124>
- CHANG, J. C., WOOTEN, E. C., TSMELZON, A., HILSENBECK, S. G., GUTIERREZ, M. C., ELLEDGE, R., MOHSIN, S., OSBORNE, C. K., CHAMNESS, G. C. et al. (2003). Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* **362** 362–369.
- CHAPELLE, O., SCHOLKOPF, B. and ZIEN, A. (2006). *Semi-Supervised Learning* (Chapelle, O. et al., Eds.); [Book Reviews]. MIT Press, Cambridge.
- CLOUGH, E. and BARRETT, T. (2016). The gene expression omnibus database. In *Statistical Genomics* 93–110. Springer, Berlin.
- CROWN, J. (2001). Docetaxel: Overview of an active drug for breast cancer. *Oncologist* **6** 1–4. https://doi.org/10.1634/theoncologist.6-suppl_3-1
- DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2019). Multi-study factor analysis. *Biometrics* **75** 337–346. [MR3953734](#) <https://doi.org/10.1111/biom.12974>
- DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2021). Bayesian multistudy factor analysis for high-throughput biological data. *Ann. Appl. Stat.* **15** 1723–1741. [MR4355073](#) <https://doi.org/10.1214/21-aos1456>
- DOMCKE, S., SINHA, R., LEVINE, D. A., SANDER, C. and SCHULTZ, N. (2013). Evaluating cell lines as tumour models by comparison of genomic profiles. *Nat. Commun.* **4** 2126. <https://doi.org/10.1038/ncomms3126>
- ELY, S. (2009). Personalized medicine: Individualized care of cancer patients. *Transl. Res.* **154** 303–308.
- FALGREEN, S., DYBKÆR, K., YOUNG, K. H., XU-MONETTE, Z. Y., EL-GALALY, T. C., LAURSEN, M. B., BØDKER, J. S., KJELDSSEN, M. K., SCHMITZ, A. et al. (2015). Predicting response to multidrug regimens in cancer patients using cell line experiments and regularised regression models. *BMC Cancer* **15** 235.
- FARNAES, L., HILDRETH, A., SWEENEY, N. M., CLARK, M. M., CHOWDHURY, S., NAHAS, S., CAKICI, J. A., BENSON, W., KAPLAN, R. H. et al. (2018). Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. *NPJ Genom. Med.* **3** 10.
- GAO, H., KORN, J. M., FERRETTI, S., MONAHAN, J. E., WANG, Y., SINGH, M., ZHANG, C., SCHNELL, C., YANG, G. et al. (2015). High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21** 1318.
- 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.
- GEELEHER, P., COX, N. J. and HUANG, R. S. (2014). Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol.* **15** R47. <https://doi.org/10.1186/gb-2014-15-3-r47>
- GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
- GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (1995). *Bayesian Data Analysis. Texts in Statistical Science Series*. CRC Press, London. [MR1385925](#)
- GEWEKE, J. (1991). *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments* **196**. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN.
- GEWEKE, J. and ZHOU, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Rev. Financ. Stud.* **9** 557–587.

- GRAY, S. L., HART, L. A., PERERA, S., SEMLA, T. P., SCHMADER, K. E. and HANLON, J. T. (2018). Meta-analysis of interventions to reduce adverse drug reactions in older adults. *J. Amer. Geriatr. Soc.* **66** 282–288. <https://doi.org/10.1111/jgs.15195>
- GRONSBELL, J. L. and CAI, T. (2018). Semi-supervised approaches to efficient evaluation of model prediction performance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 579–594. [MR3798879 https://doi.org/10.1111/rssb.12264](https://doi.org/10.1111/rssb.12264)
- GU, Z., GU, L., EILS, R., SCHLESNER, M. and BRORS, B. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics* **30** 2811–2812.
- HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.
- JIANG, Y. and WANG, M. (2010). Personalized medicine in oncology: Tailoring the right drug to the right patient. *Biomark. Med.* **4** 523–533.
- KANE, R. C., BROSS, P. F., FARRELL, A. T. and PAZDUR, R. (2003). Velcade®: US FDA approval for the treatment of multiple myeloma progressing on prior therapy. *Oncologist* **8** 508–513.
- KRÄMER, A., GREEN, J., POLLARD JR, J. and TUGENDREICH, S. (2013). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30** 523–530.
- KUMAR-SINHA, C. and CHINNAIYAN, A. M. (2018). Precision oncology in the age of integrative genomics. *Nat. Biotechnol.* **36** 46–60. <https://doi.org/10.1038/nbt.4017>
- LI, J., ZHAO, W., AKBANI, R., LIU, W., JU, Z., LING, S., VELLANO, C. P., ROEBUCK, P., YU, Q. et al. (2017). Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer Cell* **31** 225–239.
- LIU, Q., HA, M. J., BHATTACHARYYA, R., GARMIRE, L. and BALADANDAYUTHAPANI, V. (2020). Network-based matching of patients and targeted therapies for precision oncology. In *Pacific Symposium on Biocomputing* **25** 623–634. World Scientific, Singapore.
- LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14** 41–67. [MR2036762](https://doi.org/10.1214/07-BA102)
- LUCAS, J., CARVALHO, C., WANG, Q., BILD, A., NEVINS, J. R. and WEST, M. (2006). Sparse statistical modelling in gene expression genomics. *Bayesian Inference Gene Expr. Proteomics* **1** 0–1.
- MA, S., OGINO, S., PARSANA, P., NISHIHARA, R., QIAN, Z., SHEN, J., MIMA, K., MASUGI, Y., CAO, Y. et al. (2018). Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis. *Genome Biol.* **19** 1–14.
- MARTIN-ANTONIO, B., NAJJAR, A., ROBINSON, S., CHEW, C., LI, S., YVON, E., THOMAS, M., MC NIECE, I., ORLOWSKI, R. et al. (2015). Transmissible cytotoxicity of multiple myeloma cells by cord blood-derived NK cells is mediated by vesicle trafficking. *Cell Death Differ.* **22** 96.
- MCCARTHY, J. J., MCLEOD, H. L. and GINSBURG, G. S. (2013). Genomic medicine: A decade of successes, challenges, and opportunities. *Sci. Transl. Med.* **5** 189sr4. <https://doi.org/10.1126/scitranslmed.3005785>
- MULLIGAN, G., MITSIADES, C., BRYANT, B., ZHAN, F., CHNG, W. J., ROELS, S., KOENIG, E., FERGUS, A., HUANG, Y. et al. (2007). Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. *Blood* **109** 3177–3188.
- PIÑEIRO-YÁÑEZ, E., REBOIRO-JATO, M., GÓMEZ-LÓPEZ, G., PERALES-PATÓN, J., TROULÉ, K., RODRÍGUEZ, J. M., TEJERO, H., SHIMAMURA, T., LÓPEZ-CASAS, P. P. et al. (2018). PanDrugs: A novel method to prioritize anticancer drug treatments according to individual genomic data. *Gen. Med.* **10** 41.
- REULAND, S. N., GOLDSTEIN, N. B., PARTYKA, K. A., SMITH, S., LUO, Y., FUJITA, M., GONZALEZ, R., LEWIS, K., NORRIS, D. A. et al. (2012). ABT-737 synergizes with bortezomib to kill melanoma cells. *Biol. Open* **1** 92–100.
- SAHA, A., HA, M. J., ACHARYYA, S. and BALADANDAYUTHAPANI, V. (2022). Supplementary materials, and computer codes for “A Bayesian precision medicine framework for calibrating individualized therapeutic indices in cancer.” <https://doi.org/10.1214/21-AOAS1550SUPPA>, <https://doi.org/10.1214/21-AOAS1550SUPPB>
- SINHA, R., WINER, A. G., CHEVINSKY, M., JAKUBOWSKI, C., CHEN, Y.-B., DONG, Y., TICKOO, S. K., REUTER, V. E., RUSSO, P. et al. (2017). Analysis of renal cancer cell lines from two major resources enables genomics-guided cell line selection. *Nat. Commun.* **8** 15165.
- STRANSKY, N., GHANDI, M., KRYUKOV, G. V., GARRAWAY, L. A., LEHAR, J., LIU, M., SONKIN, D., KAUFFMANN, A., VENKATESAN, K. et al. (2015). Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528** 84–87.
- SUBRAMANIAN, A., NARAYAN, R., CORSELLO, S. M., PECK, D. D., NATOLI, T. E., LU, X., GOULD, J., DAVIS, J. F., TUBELLI, A. A. et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171** 1437–1452.
- TSHERNIAK, A., VAZQUEZ, F., MONTGOMERY, P. G., WEIR, B. A., KRYUKOV, G., COWLEY, G. S., GILL, S., HARRINGTON, W. F., PANTEL, S. et al. (2017). Defining a cancer dependency map. *Cell* **170** 564–576.

- WANG, H., HUANG, S., SHOU, J., SU, E. W., ONYIA, J. E., LIAO, B. and LI, S. (2006). Comparative analysis and integrative classification of NCI60 cell lines and primary tumors using gene expression profiling data. *BMC Genomics* **7** 166.
- WEINSTEIN, J. N., COLLISSE, 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. (2002). Bayesian factor regression models in the “large p , small n ” paradigm. *Bayesian Stat.* **7**.
- WILLIAMS, S. P. and MCDERMOTT, U. (2017). The pursuit of therapeutic biomarkers with high-throughput cancer cell drug screens. *Cell Chem. Biol.* **24** 1066–1074. <https://doi.org/10.1016/j.chembiol.2017.06.011>
- YANG, W., SOARES, J., GRENINGER, P., EDELMAN, E. J., LIGHTFOOT, H., FORBES, S., BINDAL, N., BEARE, D., SMITH, J. A. et al. (2012). Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41** D955–D961.
- YERLIKAYA, A., OKUR, E., ŞEKER, S. and ERIN, N. (2010). Combined effects of the proteasome inhibitor bortezomib and Hsp70 inhibitors on the B16F10 melanoma cell line. *Mol. Med. Rep.* **3** 333–339.
- ZHANG, J., BARAN, J., CROS, A., GUBERMAN, J. M., HAIDER, S., HSU, J., LIANG, Y., RIVKIN, E., WANG, J. et al. (2011). International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database* **2011**.

SEMIPARAMETRIC BAYESIAN FORECASTING OF SPATIOTEMPORAL EARTHQUAKE OCCURRENCES

BY GORDON J. ROSS^a AND ALEKSANDAR A. KOLEV^b

School of Mathematics, University of Edinburgh, ^agordon.ross@ed.ac.uk, ^baleksandar.kolev@ed.ac.uk

The Epidemic Type Aftershock Sequence (ETAS) model is a self-exciting point process which is used to model and forecast the occurrence of earthquakes in a geographical region. The ETAS model assumes that the occurrence of mainshock earthquakes follows an inhomogeneous spatial point process, with their aftershock earthquakes modelled via a separate triggering kernel. Most previous studies of the ETAS model have relied on point estimates of the model parameters, due to the complexity of the likelihood function and the difficulty in estimating an appropriate spatial mainshock distribution. In order to take estimation uncertainty into account, we instead propose a fully Bayesian formulation of the ETAS model, which uses a nonparametric Dirichlet process mixture prior to capture the spatial mainshock process, and show how efficient parameter inference can be carried out using auxiliary latent variables. We demonstrate how our model can be used for medium-term earthquake forecasts in a number of geographical regions.

REFERENCES

- ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. [MR0365969](#)
- CHIODI, M. and ADELIO, G. (2017). Mixed non-parametric and parametric estimation techniques in R package *etasFLP* for earthquakes' description. *J. Stat. Softw.* **76** 1–29.
- CHORNOBOY, E. S., SCHRAMM, L. P. and KARR, A. F. (1988). Maximum likelihood identification of neural point process systems. *Biol. Cybernet.* **59** 265–275. [MR0961117](#) <https://doi.org/10.1007/BF00332915>
- DALEY, D. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes. Volume I*. Springer, New York.
- EBRAHIMIAN, H. and JALAYER, F. (2017). Robust seismicity forecasting based on Bayesian parameter estimation for epidemiological spatio-temporal aftershock clustering models. *Sci. Rep.* **7** 1–15.
- EBRAHIMIAN, H., JALAYER, F., ASPRONE, D., LOMBARDI, A. M., MARZOCCHI, W., PROTA, A. and MANFREDI, G. (2014). Adaptive daily forecasting of seismic aftershock hazard. *Bull. Seismol. Soc. Amer.* **104** 145–161.
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. [MR0350949](#)
- 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>
- GORDON, J. S. (2017). *Nonparametric Estimation, Forecasts, and Model Evaluation of Spatial Temporal Point Process Models for California Seismicity*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—University of California, Los Angeles. [MR3697731](#)
- 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. and OAKES, D. (1974). A cluster process representation of a self-exciting process. *J. Appl. Probab.* **11** 493–503. [MR0378093](#) <https://doi.org/10.2307/3212693>
- HELMSTETTER, A. (2006). Comparison of short-term and time-independent earthquake forecast models for southern California. *Bull. Seismol. Soc. Amer.* **96** 90–106.
- HELMSTETTER, A. and SORNETTE, D. (2002). Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *J. Geophys. Res., Solid Earth* **107**.

- HOLSCHNEIDER, M., NARTEAU, C., SHEBALIN, P., PENG, Z. and SCHORLEMMER, D. (2012). Bayesian analysis of the modified Omori law. *J. Geophys. Res., Solid Earth* **117**.
- KOLEV, A. A. and ROSS, G. J. (2019). Inference for ETAS models with non-Poissonian mainshock arrival times. *Stat. Comput.* **29** 915–931. [MR3994609](#) <https://doi.org/10.1007/s11222-018-9845-z>
- KUMAZAWA, T. and OGATA, Y. (2014). Nonstationary ETAS models for nonstandard earthquakes. *Ann. Appl. Stat.* **8** 1825–1852. [MR3271355](#) <https://doi.org/10.1214/14-AOAS759>
- LIPPIELLO, E., GIACCO, F., DE ARCANGELIS, L., MARZOCCHI, W. and GODANO, C. (2014). Parameter estimation in the ETAS model: Approximations and novel methods. *Bull. Seismol. Soc. Amer.* **104** 985–994.
- LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. [MR0733519](#) <https://doi.org/10.1214/aos/1176346412>
- LOMBARDI, A. M. (2015). Estimation of the parameters of ETAS models by simulated annealing. *Sci. Rep.* **5** 8417.
- LOMBARDI, A. M. (2017). SEDA: A software package for the statistical earthquake data analysis. *Sci. Rep.* **7** 44171. <https://doi.org/10.1038/srep44171>
- MARKWICK, D. and ROSS, G. (2021). The Hierarchical Nonparametric Hawkes Process. Preprint.
- MARSAN, D. and LENGLINE, O. (2008). Extending earthquakes’ reach through cascading. *Science* **22** 1076–1079.
- 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>
- NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. [MR1823804](#) <https://doi.org/10.2307/1390653>
- OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.
- OGATA, Y. (1998). Space-time point-process models for earthquake occurrences. *Ann. Inst. Statist. Math.* **50** 379–402.
- OMI, T., OGATA, Y., HIRATA, Y. and AIHARA, K. (2016). Intermediate-term forecasting of aftershocks from an early aftershock sequence: Bayesian and ensemble forecasting approaches. *J. Geophys. Res.* **120** 2561–2578.
- RASMUSSEN, J. G. (2011). Bayesian inference for Hawkes processes. *Methodol. Comput. Appl. Probab.* **15** 623–642.
- RASMUSSEN, J. G. (2013). Bayesian inference for Hawkes processes. *Methodol. Comput. Appl. Probab.* **15** 623–642. [MR3085883](#) <https://doi.org/10.1007/s11009-011-9272-5>
- REINHART, A. and GREENHOUSE, J. (2018). Self-exciting point processes with spatial covariates: Modelling the dynamics of crime. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 1305–1329. [MR3873709](#) <https://doi.org/10.1111/rssc.12277>
- ROSS, G. J. (2021). Bayesian estimation of the ETAS model for earthquake occurrences. *Bull. Seismol. Soc. Amer.* **111** 1473–1480.
- ROSS, G. J. and KOLEV, A. A. (2022). Supplement to “Semiparametric Bayesian Forecasting of SpatioTemporal Earthquake Occurrences.” Available at <https://cran.r-project.org/web/packages/bayesianETAS/index.html>. [https://doi.org/10.1214/21-AOAS1554SUPP](#)
- ROSS, G. J. and MARKWICK, D. (2018). Dirichletprocess: An R Package for Fitting Complex Bayesian Nonparametric Models Technical report.
- SCHOENBERG, F. P. (2013). Facilitated estimation of ETAS Short note. *Bull. Seismol. Soc. Amer.* **103** 601–605.
- SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. [MR1309433](#)
- UTSU, T. and OGATA, Y. (1995). The centenary of the Omori formula for a decay law of aftershock activity. *J. Phys. Earth* **43** 1–33.
- VARGAS, N. (2012). Bayesian Point Process Modelling of Earthquake Occurrences Technical report, Ruprecht-Karls Univ. Heidelberg.
- 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, Q., SCHOENBERG, F. and JACKSON, D. (2010). Standard errors of parameter estimates in the ETAS model. *Bull. Seismol. Soc. Amer.* **106** 1989–2001.
- WIEMER, S. and WYSS, M. (2000). Minimum magnitude of completeness in earthquake catalogs: Examples from Alaska, the Western United States, and Japan. *Bull. Seismol. Soc. Amer.* **90** 859–869.
- ZHUANG, J. and MATEU, J. (2019). A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data. *J. Roy. Statist. Soc. Ser. A* **182** 919–942. [MR3955503](#)
- 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>

ESTIMATING THE STILLBIRTH RATE FOR 195 COUNTRIES USING A BAYESIAN SPARSE REGRESSION MODEL WITH TEMPORAL SMOOTHING

BY ZHENGFAN WANG^{1,a}, MIRANDA J. FIX^{2,c}, LUCIA HUG^{3,e}, ANU MISHRA^{3,f},
DANZHEN YOU^{3,g}, HANNAH BLENCOWE^{4,h}, JON WAKEFIELD^{2,d} AND
LEONTINE ALKEMA^{1,b}

¹Department of Biostatistics and Epidemiology, University of Massachusetts Amherst, ^azhengfanwang@umass.edu,
^blalkema@umass.edu

²Department of Biostatistics, University of Washington, ^cmiranda.fix@gmail.com, ^djonno@uw.edu

³Division of Data, Analytics, Planning and Monitoring, UNICEF, ^elhug@unicef.org, ^fa.mishra@imperial.ac.uk,
^gdyou@unicef.org

⁴London School of Hygiene and Tropical Medicine, University of London, ^hannah-jayne.blencowe@lshtm.ac.uk

Estimation of stillbirth rates globally is complicated because of the paucity of reliable data from countries where most stillbirths occur. We compiled data and developed a Bayesian hierarchical temporal sparse regression model for estimating stillbirth rates for 195 countries from 2000 to 2019. The model combines covariates with a temporal smoothing process so that estimates are data-driven in country-periods with high-quality data and determined by covariates for country-periods with limited or no data. Horseshoe priors are used to encourage sparseness. The model adjusts observations with alternative stillbirth definitions and accounts for various sources of uncertainty. In-sample goodness of fit and out-of-sample validation results suggest that the model is reasonably well calibrated. The model is used by the UN Interagency Group for Child Mortality Estimation to monitor the stillbirth rate for 195 countries.

REFERENCES

- AHMED, I., ALI, S. M., AMENGA-ETEGO, S., ARIFF, S., BAHL, R., BAQUI, A. H., BEGUM, N., BHANDARI, N., BHATIA, K. et al. (2018). Population-based rates, timing, and causes of maternal deaths, stillbirths, and neonatal deaths in south Asia and sub-Saharan Africa: A multi-country prospective cohort study. *The Lancet Global Health* **6** e1297–e1308. [https://doi.org/10.1016/S2214-109X\(18\)30385-1](https://doi.org/10.1016/S2214-109X(18)30385-1)
- ALKEMA, L., ZHANG, S., CHOU, D., GEMMILL, A., MOLLER, A.-B., FAT, D. M., SAY, L., MATHERS, C. and HOGAN, D. (2017). A Bayesian approach to the global estimation of maternal mortality. *Ann. Appl. Stat.* **11** 1245–1274. [MR3709559 https://doi.org/10.1214/16-AOAS1014](https://doi.org/10.1214/16-AOAS1014)
- BLENCOWE, H., COUSENS, S., JASSIR, F. B., SAY, L., CHOU, D., MATHERS, C., HOGAN, D., SHIEKH, S., QURESHI, Z. U. et al. (2016). National, regional, and worldwide estimates of stillbirth rates in 2015, with trends from 2000: A systematic analysis. *The Lancet Global Health* **4** e98–e108. [https://doi.org/10.1016/S2214-109X\(15\)00275-2](https://doi.org/10.1016/S2214-109X(15)00275-2)
- BOSE, C. L., BAUSERMAN, M., GOLDENBERG, R. L., GOUDAR, S. S., MCCLURE, E. M., PASHA, O., CARLO, W. A., GARCES, A., MOORE, J. L. et al. (2015). The global network maternal newborn health registry: A multi-national, community-based registry of pregnancy outcomes. *Reproductive Health* **12** S1. <https://doi.org/10.1186/1742-4755-12-S2-S1>
- BRADLEY, S. E. K., WINFREY, W. and CROFT, T. N. (2015). Contraceptive use and perinatal mortality in the DHS: An assessment of the quality and consistency of calendars and histories. DHS Methodological Reports No. 17. ICF International, Rockville, Maryland, USA.
- 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**. <https://doi.org/10.18637/jss.v076.i01>
- CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2009). Handling sparsity via the horseshoe. *Artificial Intelligence and Statistics* **8**.

- CURRIE, I. D. and DURBAN, M. (2002). Flexible smoothing with P -splines: A unified approach. *Stat. Model.* **2** 333–349. [MR1951589 https://doi.org/10.1191/1471082x02st039ob](https://doi.org/10.1191/1471082x02st039ob)
- 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](https://doi.org/10.1214/ss/1038425655)
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. [MR2221284 https://doi.org/10.1214/06-BA117A](https://doi.org/10.1214/06-BA117A)
- GELMAN, A. and RUBIN, B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–72.
- HUG, L., YOU, D., BLENCOWE, H., MISHRA, A., WANG, Z., FIX, M. J., WAKEFIELD, J., MORAN, A. C., GAIGBE-TOGBE, V. et al. (2021). Global, regional, and national estimates and trends in stillbirths from 2000 to 2019: A systematic assessment. *Lancet* **398** 772–785.
- KURUVILLA, S., BUSTREO, F., KUO, T., MISHRA, C., TAYLOR, K., FOGSTAD, H., GUPTA, G. R., GILMORE, K., TEMMERMAN, M. et al. (2016). The *Global strategy for women’s, children’s and adolescents’ health (2016–2030)*: A roadmap based on evidence and country experience. *Bull. World Health Organ.* **94** 398–400. <https://doi.org/10.2471/BLT.16.170431>
- LIU, L., KALTER, H. D., CHU, Y., KAZMI, N., KOFFI, A. K., AMOUZOU, A., JOOS, O., MUNOS, M. and BLACK, R. E. (2016). Understanding misclassification between neonatal deaths and stillbirths: Empirical evidence from Malawi. *PLoS ONE* **11** e0168743. <https://doi.org/10.1371/journal.pone.0168743>
- PEDERSEN, J. and LIU, J. (2012). Child mortality estimation: Appropriate time periods for child mortality estimates from full birth histories. *PLoS Med.* **9** e1001289. <https://doi.org/10.1371/journal.pmed.1001289>
- PIIRONEN, J. and VEHTARI, A. (2017a). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Stat.* **11** 5018–5051. [MR3738204 https://doi.org/10.1214/17-EJS1337SI](https://doi.org/10.1214/17-EJS1337SI)
- PIIRONEN, J. and VEHTARI, A. (2017b). Comparison of Bayesian predictive methods for model selection. *Stat. Comput.* **27** 711–735. [MR3613594 https://doi.org/10.1007/s11222-016-9649-y](https://doi.org/10.1007/s11222-016-9649-y)
- STANTON, C., LAWN, J. E., RAHMAN, H., WILCZYNSKA-KETENDE, K. and HILL, K. (2006). Stillbirth rates: Delivering estimates in 190 countries. *Lancet* **367** 1487–1494. [https://doi.org/10.1016/S0140-6736\(06\)68586-3](https://doi.org/10.1016/S0140-6736(06)68586-3)
- STAN DEVELOPMENT TEAM (2018). RStan: The R interface to Stan. R package version 2.18.2.
- UN INTER-AGENCY GROUP FOR CHILD MORTALITY ESTIMATION (2020). A neglected tragedy: The global burden of stillbirths. Report of the UN Inter-agency Group for Child Mortality Estimation, 2020. United Nations Children’s Fund, New York.
- VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27** 1413–1432. [MR3647105 https://doi.org/10.1007/s11222-016-9696-4](https://doi.org/10.1007/s11222-016-9696-4)
- VEHTARI, A., GABRY, J., MAGNUSSON, M., YAO, Y. and GELMAN, A. (2019). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.2.0.
- VEHTARI, A., GELMAN, A., SIMPSON, D., CARPENTER, B. and BÜRKNER, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Anal.* **16** 667–718. Includes comments and discussions by seven discussants and a rejoinder by the authors. [MR4298989 https://doi.org/10.1214/20-ba1221](https://doi.org/10.1214/20-ba1221)
- WANG, Z., FIX, M. J., HUG, L., MISHRA, A., YOU, D., BLENCOWE, H., WAKEFIELD, J. and ALKEMA, L. (2022). Supplement to “Estimating the stillbirth rate for 195 countries using a Bayesian sparse regression model with temporal smoothing.” <https://doi.org/10.1214/21-AOAS1571SUPP>
- WHO DEPARTMENT OF INFORMATION, EVIDENCE, AND RESEARCH (2018). WHO methods and data sources for life tables 1990–2016.
- WOODS, R. (2008). Long-term trends in fetal mortality: Implications for developing countries. *Bull. World Health Organ.* **2008** 460–466. <https://doi.org/10.2471/BLT.07.043471>
- WORLD HEALTH ORGANIZATION (2014). *Every Newborn: An Action Plan to End Preventable Deaths*. World Health Organization, Geneva. OCLC: 918973731.
- WORLD HEALTH ORGANIZATION (2019). *International Classification of Disease (ICD), Revised Definition for ICD-11*.

FUNCTIONAL RANDOM EFFECTS MODELING OF BRAIN SHAPE AND CONNECTIVITY

BY EARDI LILA^{1,a} AND JOHN A. D. ASTON^{2,b}

¹Department of Biostatistics, University of Washington, ^aelila@uw.edu

²Statistical Laboratory, University of Cambridge, ^bj.aston@statslab.cam.ac.uk

We present a statistical framework that jointly models brain shape and functional connectivity which are two complex aspects of the brain that have been classically studied independently. We adopt a Riemannian modeling approach to account for the non-Euclidean geometry of the space of shapes and the space of connectivity that constrains trajectories of covariation to be valid statistical estimates. In order to disentangle genetic sources of variability from those driven by unique environmental factors, we embed a functional random effects model in the Riemannian framework. We apply the proposed model to the Human Connectome Project dataset to explore spontaneous covariation between brain shape and connectivity in young healthy individuals.

REFERENCES

- ALMASY, L., DYER, T. D. and BLANGERO, J. (1997). Bivariate quantitative trait linkage analysis: Pleiotropy versus co-incident linkages. In *Genetic Epidemiology* **14** 953–958.
- AMOS, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* **54** 535–543.
- ARGUILLÈRE, S., MILLER, M. I. and YOUNES, L. (2016). Diffeomorphic surface registration with atrophy constraints. *SIAM J. Imaging Sci.* **9** 975–1003. [MR3521541](#) <https://doi.org/10.1137/15M104431X>
- ARSIGNY, V., FILLARD, P., PENNEC, X. and AYACHE, N. (2006). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* **29** 328–347. [MR2288028](#) <https://doi.org/10.1137/050637996>
- ARSLAN, S., KTEA, S. I., MAKROPOULOS, A., ROBINSON, E. C., RUECKERT, D. and PARISOT, S. (2018). Human brain mapping: A systematic comparison of parcellation methods for the human cerebral cortex. *NeuroImage* **170** 5–30. <https://doi.org/10.1016/j.neuroimage.2017.04.014>
- AYACHIT, U. (2015). *The ParaView Guide: A Parallel Visualization Application*. Kitware, Inc., Clifton Park, NY, USA.
- BARBER, A. D., HEGARTY, C. E., LINDQUIST, M. and KARLSGODT, K. H. (2021). Heritability of functional connectivity in resting state: Assessment of the dynamic mean, dynamic variance, and static connectivity across networks. *Cereb. Cortex* **31** 2834–2844.
- BATES, D., MÄCHLER, M., BOLKER, B. M. and WALKER, S. C. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67** 1–48.
- BULLMORE, E. and SPORNS, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10** 186–198.
- CALHOUN, V. D., LIU, J. and ADALI, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage* **45** S163–S172.
- CARMICHAEL, I., CALHOUN, B. C., HOADLEY, K. A. et al. (2021). Joint and individual analysis of breast cancer histologic images and genomic covariates. *Ann. Appl. Stat.* **15** 1697–1722. [MR4355072](#) <https://doi.org/10.1214/20-aoas1433>
- CHARLIER, B., CHARON, N. and TROUVÉ, A. (2017). The Fshape framework for the variability analysis of functional shapes. *Found. Comput. Math.* **17** 287–357. [MR3627451](#) <https://doi.org/10.1007/s10208-015-9288-2>
- CHARLIER, B., NARDI, G. and TROUVÉ, A. (2015). The matching problem between functional shapes via a BV-penalty term: A Γ -convergence result. 1–31.
- CHARON, N. and TROUVÉ, A. (2014). Functional currents: A new mathematical tool to model and analyse functional shapes. *J. Math. Imaging Vision* **48** 413–431.

- CHEN, H. and WANG, Y. (2008). A penalized spline approach to functional mixed effects model analysis. *Biometrics* **64** 751–761.
- DAHL, A., IOTCHKOVA, V., BAUD, A., JOHANSSON, Å., GYLLENSTEN, U., SORANZO, N., MOTT, R., KRANIS, A. and MARCHINI, J. (2016). A multiple-phenotype imputation method for genetic studies. *Nat. Genet.* **48** 466–472.
- DESIKAN, R. S., SÉGONNE, F., FISCHL, B., QUINN, B. T., DICKERSON, B. C., BLACKER, D., BUCKNER, R. L., DALE, A. M., MAGUIRE, R. P. et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31** 968–980.
- DRYDEN, I. L., KOLOYDENKO, A. and ZHOU, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann. Appl. Stat.* **3** 1102–1123. [MR2750388 https://doi.org/10.1214/09-AOAS249](https://doi.org/10.1214/09-AOAS249)
- 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](https://doi.org/10.1002/9781119072492)
- ELTZNER, B., HUCKEMANN, S. and MARDIA, K. V. (2018). Torus principal component analysis with applications to RNA structure. *Ann. Appl. Stat.* **12** 1332–1359. [MR3834306 https://doi.org/10.1214/17-AOAS1115](https://doi.org/10.1214/17-AOAS1115)
- FENG, Q., JIANG, M., HANNIG, J. and MARRON, J. S. (2018). Angle-based joint and individual variation explained. *J. Multivariate Anal.* **166** 241–265. [MR3799646 https://doi.org/10.1016/j.jmva.2018.03.008](https://doi.org/10.1016/j.jmva.2018.03.008)
- FEYDY, J., CHARLIER, B., VIALARD, F. X. and PEYRÉ, G. (2017). Optimal transport for diffeomorphic registration. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2017. MICCAI 2017. Lecture Notes in Computer Science* (M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. Collins and S. Duchesne, eds.) **10433** 291–299. Springer, Cham.
- FIECAS, M., CRIBBEN, I., BAHKTIARI, R. and CUMMINE, J. (2017). A variance components model for statistical inference on functional connectivity networks. *NeuroImage* **149** 256–266. <https://doi.org/10.1016/j.neuroimage.2017.01.051>
- FISCHL, B., SERENO, M. I. and DALE, A. M. (1999). Cortical surface-based analysis II: Inflation, flattening, and a surface-based coordinate system. *NeuroImage* **9** 195–207.
- FISCHL, B., VAN DER KOUWE, A., DESTRIEUX, C., HALGREN, E., SÉGONNE, F., SALAT, D. H., BUSA, E., SEIDMAN, L. J., GOLDSTEIN, J. et al. (2004). Automatically parcellating the human cerebral cortex. *Cereb. Cortex* **14** 11–22. <https://doi.org/10.1093/cercor/bhg087>
- FRANCO, A. R., LING, J., CAPRIHAN, A., CALHOUN, V. D., JUNG, R. E., HEILEMAN, G. L. and MAYER, A. R. (2008). Multimodal and multi-tissue measures of connectivity revealed by joint independent component analysis. *IEEE J. Sel. Top. Signal Process.* **2** 986–997.
- GILMORE, J. H., KNICKMEYER, R. C. and GAO, W. (2018). Imaging structural and functional brain development in early childhood. *Nat. Rev. Neurosci.* **19** 123–137.
- GLASSER, M. F., SOTIROPOULOS, S. N., WILSON, J. A., COALSON, T. S., FISCHL, B., ANDERSSON, J. L., XU, J., JBABDI, S., WEBSTER, M. et al. (2013). The minimal preprocessing pipelines for the human connectome project. *NeuroImage* **80** 105–124.
- GLASSER, M. F., SMITH, S. M., MARCUS, D. S., ANDERSSON, J. L. R., AUERBACH, E. J., BEHRENS, T. E. J., COALSON, T. S., HARMS, M. P., JENKINSON, M. et al. (2016). The human connectome project's neuroimaging approach. *Nat. Neurosci.* **19** 1175–1187.
- GORDON, E. M., LAUMANN, T. O., ADEYEMO, B., HUCKINS, J. F., KELLEY, W. M. and PETERSEN, S. E. (2016). Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cereb. Cortex* **26** 288–303. <https://doi.org/10.1093/cercor/bhu239>
- GUO, W. (2002). Functional mixed effects models. *Biometrics* **58** 121–128. [MR1891050 https://doi.org/10.1111/j.0006-341X.2002.00121.x](https://doi.org/10.1111/j.0006-341X.2002.00121.x)
- HAZLETT, H. C., GU, H., MUNSELL, B. C., KIM, S. H., STYNER, M., WOLFF, J. J., ELISON, J. T., SWANSON, M. R., ZHU, H. et al. (2017). Early brain development in infants at high risk for autism spectrum disorder. *Nature* **542** 348–351.
- IM, K., LEE, J. M., LYTTELTON, O., KIM, S. H., EVANS, A. C. and KIM, S. I. (2008). Brain size and cortical structure in the adult human brain. *Cereb. Cortex* **18** 2181–2191.
- JERMYN, I. H., KURTEK, S., KLASSEN, E. and SRIVASTAVA, A. (2012). Elastic shape matching of parameterized surfaces using square root normal fields. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **7576 LNCS** 804–817.
- JERMYN, I. H., KURTEK, S., LAGA, H. and SRIVASTAVA, A. (2017). Elastic shape analysis of three-dimensional objects. *Synth. Lect. Comput. Vis.* **7** 1–185.
- JOE, H. (2006). Generating random correlation matrices based on partial correlations. *J. Multivariate Anal.* **97** 2177–2189. [MR2301633 https://doi.org/10.1016/j.jmva.2005.05.010](https://doi.org/10.1016/j.jmva.2005.05.010)
- KANG, H. M., SUL, J. H., SERVICE, S. K., ZAITLEN, N. A., KONG, S. Y., FREIMER, N. B., SABATTI, C. and ESKIN, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42** 348–354.

- KIRKPATRICK, M. and HECKMAN, N. (1989). A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *J. Math. Biol.* **27** 429–450. [MR1009899](#) <https://doi.org/10.1007/BF00290638>
- KURTEK, S., KLASSEN, E., DING, Z., JACOBSON, S. W., JACOBSON, J. L., AVISON, M. J. and SRIVASTAVA, A. (2011). Parameterization-invariant shape comparisons of anatomical surfaces. *IEEE Trans. Med. Imag.* **30** 849–858.
- KURTEK, S., SRIVASTAVA, A., KLASSEN, E. and DING, Z. (2012). Statistical modeling of curves using shapes and related features. *J. Amer. Statist. Assoc.* **107** 1152–1165. [MR3010902](#) <https://doi.org/10.1080/01621459.2012.699770>
- LANGE, K. (2002). *Mathematical and Statistical Methods for Genetic Analysis*, 2nd ed. *Statistics for Biology and Health*. Springer, New York. [MR1892279](#) <https://doi.org/10.1007/978-0-387-21750-5>
- LENROOT, R. K. and GIEDD, J. N. (2008). The changing impact of genes and environment on brain development during childhood and adolescence: Initial findings from a neuroimaging study of pediatric twins. *Dev. Psychopathol.* **20** 1161–1175.
- LEWIS, A. S. and OVERTON, M. L. (2013). Nonsmooth optimization via quasi-Newton methods. *Math. Program.* **141** 135–163. [MR3097282](#) <https://doi.org/10.1007/s10107-012-0514-2>
- LILA, E. and ASTON, J. A. D. (2020). Statistical analysis of functions on surfaces, with an application to medical imaging. *J. Amer. Statist. Assoc.* **115** 1420–1434. [MR4143475](#) <https://doi.org/10.1080/01621459.2019.1635479>
- LILA, E. and ASTON, J. A. (2022). Supplement to “Functional random effects modeling of brain shape and connectivity.” <https://doi.org/10.1214/21-AOAS1572SUPP>
- LIU, B., WANG, L. and CAO, J. (2017). Estimating functional linear mixed-effects regression models. *Comput. Statist. Data Anal.* **106** 153–164. [MR3566081](#) <https://doi.org/10.1016/j.csda.2016.09.009>
- LUO, S., SONG, R., STYNER, M., GILMORE, J. H. and ZHU, H. (2019). FSEM: Functional structural equation models for twin functional data. *J. Amer. Statist. Assoc.* **114** 344–357. [MR3941259](#) <https://doi.org/10.1080/01621459.2017.1407773>
- MARDIA, K. V. and PATRANGENARU, V. (2005). Directions and projective shapes. *Ann. Statist.* **33** 1666–1699. [MR2166559](#) <https://doi.org/10.1214/009053605000000273>
- MARRON, J. S. and ALONSO, A. M. (2014). Overview of object oriented data analysis. *Biom. J.* **56** 732–753. [MR3258083](#) <https://doi.org/10.1002/bimj.201300072>
- MILLER, M. I., TROUVÉ, A. and YOUNES, L. (2006). Geodesic shooting for computational anatomy. *J. Math. Imaging Vision* **24** 209–228. [MR2227097](#) <https://doi.org/10.1007/s10851-005-3624-0>
- MORRIS, J. S. and CARROLL, R. J. (2006). Wavelet-based functional mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 179–199. [MR2188981](#) <https://doi.org/10.1111/j.1467-9868.2006.00539.x>
- PENNEC, X., FILLARD, P. and AYACHE, N. (2006). A Riemannian framework for tensor computing. *Int. J. Comput. Vis.* **66** 41–66.
- PIGOLI, D., ASTON, J. A. D., DRYDEN, I. L. and SECCHI, P. (2014). Distances and inference for covariance operators. *Biometrika* **101** 409–422. [MR3215356](#) <https://doi.org/10.1093/biomet/asu008>
- PIZER, S. M., JUNG, S., GOSWAMI, D., VICORY, J., ZHAO, X., CHAUDHURI, R., DAMON, J. N., HUCKEMANN, S. and MARRON, J. S. (2013). Nested sphere statistics of skeletal models. In *Innovations for Shape Analysis. Math. Vis.* 93–115. Springer, Heidelberg. [MR3075829](#) https://doi.org/10.1007/978-3-642-34141-0_5
- PIZZAGALLI, F., AUZIAS, G., YANG, Q., MATHIAS, S. R., FASKOWITZ, J., BOYD, J. D., AMINI, A., RIVIÈRE, D., MCMAHON, K. L. et al. (2020). The reliability and heritability of cortical folds and their genetic correlations across hemispheres. *Commun. Biol.* **3**.
- POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M. et al. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.
- QIN, L. (2005). Functional mixed-effects model for periodic data. *Biostatistics* **7** 225–234.
- RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR2168993](#)
- REIMHERR, M. and NICOLAE, D. (2016). Estimating variance components in functional linear models with applications to genetic heritability. *J. Amer. Statist. Assoc.* **111** 407–422. [MR3494668](#) <https://doi.org/10.1080/01621459.2015.1016224>
- RISK, B. B. and ZHU, H. (2021). ACE of space: Estimating genetic components of high-dimensional imaging data. *Biostatistics* **22** 131–147. [MR4206618](#) <https://doi.org/10.1093/biostatistics/kxz022>
- ROBINSON, E. C., JBABDI, S., GLASSER, M. F., ANDERSSON, J., BURGESS, G. C., HARMS, M. P., SMITH, S. M., VAN ESSEN, D. C. and JENKINSON, M. (2014). MSM: A new flexible framework for multimodal surface matching. *NeuroImage* **100** 414–426.

- ROBINSON, E. C., GARCIA, K., GLASSER, M. F., CHEN, Z., COALSON, T. S., MAKROPOULOS, A., BOZEK, J., WRIGHT, R., SCHUH, A. et al. (2018). Multimodal surface matching with higher-order smoothness constraints. *NeuroImage* **167** 453–465.
- SCHEIPL, F., STAIKU, A.-M. and GREVEN, S. (2015). Functional additive mixed models. *J. Comput. Graph. Statist.* **24** 477–501. [MR3357391 https://doi.org/10.1080/10618600.2014.901914](https://doi.org/10.1080/10618600.2014.901914)
- 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. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **45** 151.
- SMITH, S. M., BECKMANN, C. F., ANDERSSON, J., AUERBACH, E. J., BIJSTERBOSCH, J., DOUAUD, G., DUFF, E., FEINBERG, D. A., GRIFFANTI, L. et al. (2013). Resting-state fMRI in the human connectome project. *NeuroImage* **80** 144–168.
- SMITH, S. M., NICHOLS, T. E., VIDAURRE, D., WINKLER, A. M., BEHRENS, T. E. J., GLASSER, M. F., UGURBIL, K., BARCH, D. M., VAN ESSEN, D. C. et al. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nat. Neurosci.* **18** 1565–1567.
- SU, J., KURTEK, S., KLASSEN, E. and SRIVASTAVA, A. (2014). Statistical analysis of trajectories on Riemannian manifolds: Bird migration, hurricane tracking and video surveillance. *Ann. Appl. Stat.* **8** 530–552. [MR3192001 https://doi.org/10.1214/13-AOAS701](https://doi.org/10.1214/13-AOAS701)
- SUDLOW, C., GALLACHER, J., ALLEN, N., BERAL, V., BURTON, P., DANESH, J., DOWNEY, P., ELLIOTT, P., GREEN, J. et al. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12** 1001779.
- SUI, J., PEARLSON, G., CAPRIHAN, A., ADALI, T., KIEHL, K. A., LIU, J., YAMAMOTO, J. and CALHOUN, V. D. (2011). Discriminating schizophrenia and bipolar disorder by fusing fMRI and DTI in a multimodal CCA+ joint ICA model. *NeuroImage* **57** 839–855.
- VAILLANT, M., MILLER, M. I., YOUNES, L. and TROUVÉ, A. (2004). Statistics on diffeomorphisms via tangent space representations. *NeuroImage* **23** S161–S169.
- VAN ESSEN, D. C., GLASSER, M. F., DIERKER, D. L., HARWELL, J. and COALSON, T. (2012). Parcellations and hemispheric asymmetries of human cerebral cortex analyzed on surface-based atlases. *Cereb. Cortex* **22** 2241–2262.
- VAN ESSEN, D. C., SMITH, S. M., BARCH, D. M., BEHRENS, T. E. J., YACOB, E. and UGURBIL, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage* **80** 62–79.
- WANG, B., SVERDLOV, S. and THOMPSON, E. (2017). Efficient estimation of realized kinship from single nucleotide polymorphism genotypes. *Genetics* **205** 1063–1078. <https://doi.org/10.1534/genetics.116.197004>
- WIG, G. S., LAUMANN, T. O. and PETERSEN, S. E. (2014). An approach for parcellating human cortical areas using resting-state correlations. *NeuroImage* **93** 276–291.
- WU, H. and ZHANG, J.-T. (2002). Local polynomial mixed-effects models for longitudinal data. *J. Amer. Statist. Assoc.* **97** 883–897. [MR1941417 https://doi.org/10.1198/016214502388618672](https://doi.org/10.1198/016214502388618672)
- XIA, C. H., MA, Z., CIRIC, R., GU, S., BETZEL, R. F., KACZKURKIN, A. N., CALKINS, M. E., COOK, P. A., GARCÍA DE LA GARZA, A. et al. (2018). Linked dimensions of psychopathology and connectivity in functional brain networks. *Nat. Commun.* **9**.
- XUE, W., DUBOIS BOWMAN, F., PILEGGI, A. V. and MAYER, A. R. (2015). A multimodal approach for determining brain networks by jointly modeling functional and structural connectivity. *Front. Comput. Neurosci.* **9** 1–11.
- YEO, B. T. T., SABUNCU, M. R., VERCAUTEREN, T., AYACHE, N., FISCHL, B. and GOLLAND, P. (2010). Spherical demons: Fast diffeomorphic landmark-free surface registration. *IEEE Trans. Med. Imag.* **29** 650–668.
- YEO, B. T. T., KRIENEN, F. M., SEPULCRE, J., SABUNCU, M. R., LASHKARI, D., HOLLINSHEAD, M., ROFFMAN, J. L., SMOLLER, J. W., ZÖLLEI, L. et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106** 1125–1165.
- YOUNES, L. (2010). *Shapes and Diffeomorphisms. Applied Mathematical Sciences*. Springer, Berlin, Heidelberg.
- ZHANG, Z., ALLEN, G. I., ZHU, H. and DUNSON, D. (2019). Tensor network factorizations: Relationships between brain structural connectomes and traits. *NeuroImage* **197** 330–343.
- ZHOU, L., HUANG, J. Z. and CARROLL, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika* **95** 601–619. [MR2443178 https://doi.org/10.1093/biomet/asn035](https://doi.org/10.1093/biomet/asn035)
- ZHOU, X. and STEPHENS, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11** 407–409.
- ZITOVÁ, B. and FLUSSER, J. (2003). Image registration methods: A survey. *Image Vis. Comput.* **21** 977–1000.

HIERARCHICAL RESAMPLING FOR BAGGING IN MULTISTUDY PREDICTION WITH APPLICATIONS TO HUMAN NEUROCHEMICAL SENSING

BY GABRIEL LOEWINGER^{1,a}, PRASAD PATIL^{2,b}, KENNETH T. KISHIDA^{3,c} AND GIOVANNI PARMIGIANI^{4,d}

¹Department of Biostatistics, Harvard T.H. Chan School of Public Health, ^agloewinger@gmail.com

²Department of Biostatistics, Boston University School of Public Health, ^bpatil@bu.edu

³Department of Physiology and Pharmacology, Wake Forest School of Medicine, ^ckkishida@wakehealth.edu

⁴Department of Data Science, Dana Farber Cancer Institute, ^dgp@jimmy.harvard.edu

We propose the “study strap ensemble,” which combines advantages of two common approaches to fitting prediction models when multiple training datasets (“studies”) are available: pooling studies and fitting one model vs. averaging predictions from multiple models each fit to individual studies. The study strap ensemble fits models to bootstrapped datasets or “pseudo-studies.” These are generated by resampling from multiple studies with a hierarchical resampling scheme that generalizes the randomized cluster bootstrap. The study strap is controlled by a tuning parameter that determines the proportion of observations to draw from each study. When the parameter is set to its lowest value, each pseudo-study is resampled from only a single study. When it is high, the study strap ignores the multistudy structure and generates pseudo-studies by merging the datasets and drawing observations like a standard bootstrap. We empirically show the optimal tuning value often lies in between and prove that special cases of the study strap draw the merged dataset and the set of original studies as pseudo-studies. We extend the study strap approach with an ensemble weighting scheme that utilizes information in the distribution of the covariates of the test dataset.

Our work is motivated by neuroscience experiments using real-time neurochemical sensing during awake behavior in humans. Current techniques to perform this kind of research require measurements from an electrode placed in the brain during awake neurosurgery and rely on prediction models to estimate neurotransmitter concentrations from the electrical measurements recorded by the electrode. These models are trained by combining multiple datasets that are collected *in vitro* under heterogeneous conditions in order to promote accuracy of the models when applied to data collected in the brain. A prevailing challenge is deciding how to combine studies or ensemble models trained on different studies to enhance model generalizability.

Our methods produce marked improvements in simulations and in this application. All methods are available in the `studyStrap` CRAN package.

REFERENCES

- BANG, D., KISHIDA, K. T., LOHRENZ, T., WHITE, J. P., LAXTON, A. W., TATTER, S. B., FLEMING, S. M. and MONTAGUE, P. R. (2020). Sub-second dopamine and serotonin signaling in human striatum during perceptual decision-making. *Neuron* **108** 999–1010.
- BICKEL, S., BRÜCKNER, M. and SCHEFFER, T. (2009). Discriminative learning under covariate shift. *J. Mach. Learn. Res.* **10** 2137–2155. [MR2550104](#)
- BREIMAN, L. (1996a). Bagging predictors. *Mach. Learn.* **24** 123–140.
- BREIMAN, L. (1996b). Stacked regressions. *Mach. Learn.* **24** 49–64.
- DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics **1**. Cambridge Univ. Press, Cambridge. [MR1478673](#)
<https://doi.org/10.1017/CBO9780511802843>

- DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2019). Multi-study factor analysis. *Biometrics* **75** 337–346. [MR3953734 https://doi.org/10.1111/biom.12974](https://doi.org/10.1111/biom.12974)
- FARAHANI, A., VOGHOEI, S., RASHEED, K. and ARABNIA, H. R. (2020). A brief review of domain adaptation. Available at [arXiv:2010.03978](https://arxiv.org/abs/2010.03978).
- GLOCKER, B., ROBINSON, R., CASTRO, D. C., DOU, Q. and KONUKOGLU, E. (2019). Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. Available at [arXiv:1910.04597](https://arxiv.org/abs/1910.04597).
- GONG, B., SHI, Y., SHA, F. and GRAUMAN, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* 2066–2073.
- GUAN, Z., PARMIGIANI, G. and PATIL, P. (2020). Merging vs. ensembling in multistudy machine learning: Theoretical insight from random effects. Available at [arXiv:1905.07382](https://arxiv.org/abs/1905.07382).
- GUO, J., SHAH, D. and BARZILAY, R. (2018). Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* 4694–4703. Association for Computational Linguistics, Brussels, Belgium.
- GUO, W., WANG, S., DING, P., WANG, Y. and JORDAN, M. I. (2021). Multi-source causal inference using control variates. Available at [arXiv:2103.16689](https://arxiv.org/abs/2103.16689).
- HU, S., ZHANG, K., CHEN, Z. and CHAN, L. (2020). Domain generalization via multidomain discriminant analysis. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference* (R. P. Adams and V. Gogate, eds.). *Proceedings of Machine Learning Research* **115** 292–302. PMLR.
- HUANG, J., GRETTON, A., BORGWARDT, K., SCHÖLKOPF, B. and SMOLA, A. (2007). Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems* (B. Schölkopf, J. Platt and T. Hoffman, eds.) **19**. MIT Press.
- JOHNSON, J. A., RODEBERG, N. T. and WIGHTMAN, R. M. (2016). Failure of standard training sets in the analysis of fast-scan cyclic voltammetry data. *ACS Chemical Neuroscience* **7** 349–359.
- KEITHLEY, R. and WIGHTMAN, M. (2011). Assessing principal component regression prediction of neurochemicals detected with fast-scan cyclic voltammetry. *ACS Chemical Neuroscience* **2** 514–525.
- KISHIDA, K. T., SAEZ, I., LOHRENZ, T., WITCHER, M. R., LAXTON, A. W., TATTER, S. B., WHITE, J. P., ELLIS, T. L., PHILLIPS, P. E. M. et al. (2016). Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *Proc. Natl. Acad. Sci. USA* **113** 200–205.
- KOUW, W. M. and LOOG, M. (2019). An introduction to domain adaptation and transfer learning. Available at [arXiv:1812.11806](https://arxiv.org/abs/1812.11806).
- KRUSKAL, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29** 1–27. [MR0169712 https://doi.org/10.1007/BF02289565](https://doi.org/10.1007/BF02289565)
- KRUSKAL, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **29** 115–129. [MR0169713 https://doi.org/10.1007/BF02289694](https://doi.org/10.1007/BF02289694)
- LOEWINGER, G., PATIL, P., KISHIDA, K. T. and PARMIGIANI, G. (2022). Supplement to “Hierarchical resampling for bagging in multistudy prediction with applications to human neurochemical sensing.” <https://doi.org/10.1214/21-AOAS1574SUPP>
- MANSOUR, Y., MOHRI, M. and ROSTAMIZADEH, A. (2009). Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems* (D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds.) **21**. Curran Associates, Red Hook.
- MONTAGUE, P. R. and KISHIDA, K. T. (2018). Computational underpinnings of neuromodulation in humans. *Cold Spring Harbor Symp. Quant. Biol.* **83** 71–82. <https://doi.org/10.1101/sqb.2018.83.038166>
- MORAN, R. J., KISHIDA, K. T., LOHRENZ, T., SAEZ, I., LAXTON, A. W., WITCHER, M. R., TATTER, S. B., ELLIS, T. L., PHILLIPS, P. E. et al. (2018). The protective action encoding of serotonin transients in the human brain. *Neuropsychopharmacology* **43** 1425–1435. <https://doi.org/10.1038/npp.2017.304>
- PATIL, P. and PARMIGIANI, G. (2018). Training replicable predictors in multiple studies. *Proc. Natl. Acad. Sci. USA* **115** 2578–2583. [MR3779784 https://doi.org/10.1073/pnas.1708283115](https://doi.org/10.1073/pnas.1708283115)
- POWERS, S., HASTIE, T. and TIBSHIRANI, R. (2015). Customized training with an application to mass spectrometric imaging of cancer tissue. *Ann. Appl. Stat.* **9** 1709–1725. [MR3456351 https://doi.org/10.1214/15-AOAS866](https://doi.org/10.1214/15-AOAS866)
- RAMCHANDRAN, M., PATIL, P. and PARMIGIANI, G. (2020). Tree-weighting for multistudy ensemble learners. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* **25** 451–462.
- RAMON, E., BELANCHE-MUÑOZ, L. and PÉREZ-ENCISO, M. (2019). HIV drug resistance prediction with weighted categorical kernel functions. *BMC Bioinform.* **20** 410. <https://doi.org/10.1186/s12859-019-2991-2>
- RASHID, N. U., LI, Q., YEH, J. J. and IBRAHIM, J. G. (2020). Modeling between-study heterogeneity for improved replicability in gene signature selection and clinical prediction. *J. Amer. Statist. Assoc.* **115** 1125–1138. [MR4143454 https://doi.org/10.1080/01621459.2019.1671197](https://doi.org/10.1080/01621459.2019.1671197)
- REN, B., PATIL, P., DOMINICI, F., PARMIGIANI, G. and TRIPPA, L. (2021). Cross-study learning for generalist and specialist predictions. Available at [arXiv:2007.12807](https://arxiv.org/abs/2007.12807).

- RODEBERG, N. T., SANDBERG, S. G., JOHNSON, J. A., PHILLIPS, P. E. M. and WIGHTMAN, R. M. (2017). Hitchhiker’s guide to voltammetry: Acute and chronic electrodes for in vivo fast-scan cyclic voltammetry. *ACS Chemical Neuroscience* **8** 221–234.
- ROY, A., LAVINE, I., HERRING, A. H. and DUNSON, D. B. (2021). Perturbed factor analysis: Accounting for group differences in exposure profiles. *Ann. Appl. Stat.* **15** 1386–1404. [MR4316654](#) <https://doi.org/10.1214/20-aos1435>
- SAMMON, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* **C-18** 401–409.
- SHEPARD, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika* **27** 125–140. [MR0140376](#) <https://doi.org/10.1007/BF02289630>
- SHIMODAIRA, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference* **90** 227–244. [MR1795598](#) [https://doi.org/10.1016/S0378-3758\(00\)00115-4](https://doi.org/10.1016/S0378-3758(00)00115-4)
- SUGIYAMA, M., NAKAJIMA, S., KASHIMA, H., BUENAU, P. and KAWANABE, M. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems* (J. Platt, D. Koller, Y. Singer and S. Roweis, eds.) **20**. Curran Associates.
- SUN, S., SHI, H. and WU, Y. (2015). A survey of multi-source domain adaptation. *Inf. Fusion* **24** 84–92.
- SUN, Q., CHATTOPADHYAY, R., PANCHANATHAN, S. and YE, J. (2011). A two-stage weighting framework for multi-source domain adaptation. In *Advances in Neural Information Processing Systems* (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Q. Weinberger, eds.) **24**. Curran Associates, Red Hook.
- VENTZ, S., MAZUMDER, R. and TRIPPA, L. (2021). Integration of survival data from multiple studies. *Biometrics*.
- VOLKOW, N., WISE, R. and BALER, R. (2017). The dopamine motive system: Implications for drug and food addiction. *Nat. Rev. Neurosci.* **18** 741–752.
- WANG, J., LAN, C., LIU, C., OUYANG, Y. and QIN, T. (2021). Generalizing to unseen domains: A survey on domain generalization. Available at [arXiv:2103.03097](#).
- XU, X., ZHOU, X., VENKATESAN, R., SWAMINATHAN, G. and MAJUMDER, O. (2019). *d*-SNE: Domain adaptation using stochastic neighborhood embedding. Available at [arXiv:1905.12775](#).
- YANG, Z., AL-DAHIDI, S., BARALDI, P., ZIO, E. and MONTELATICI, L. (2020). A novel concept drift detection method for incremental learning in nonstationary environments. *IEEE Trans. Neural Netw. Learn. Syst.* **31** 309–320.
- ZHANG, Y. and YANG, Q. (2021). A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.* 1–1.
- ZHANG, Y., PATIL, P., JOHNSON, W. E. and PARMIGIANI, G. (2020). Robustifying genomic classifiers to batch effects via ensemble learning. *Bioinformatics*.

NETWORK DIFFERENTIAL CONNECTIVITY ANALYSIS

BY SEN ZHAO^{1,a} AND ALI SHOJAIE^{2,b}

¹Google Research, Google LLC, ^asenzhaog@google.com

²Department of Biostatistics, University of Washington, ^bashojaie@uw.edu

Identifying differences in networks has become a canonical problem in many biological applications. Existing methods try to accomplish this goal by either directly comparing the estimated structures of two networks or testing the null hypothesis that the covariance or inverse covariance matrices in two populations are identical. However, estimation approaches do not provide measures of uncertainty, for example, p -values, whereas existing testing approaches could lead to misleading results, as we illustrate in this paper. To address these shortcomings, we propose a *qualitative* hypothesis testing framework which tests whether the connectivity *structures* in the two networks are the same. Our framework is especially appropriate if the goal is to identify nodes or edges that are differentially connected. No existing approach could test such hypotheses and provide corresponding measures of uncertainty. Theoretically, we show that, under appropriate conditions, our proposal correctly controls the type-I error rate in testing the qualitative hypothesis. Empirically, we demonstrate the performance of our proposal using simulation studies and applications in cancer genomics.

REFERENCES

- BARABÁSI, A.-L., GULBAHCE, N. and LOSCALZO, J. (2011). Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12** 56–68. <https://doi.org/10.1038/nrg2918>
- BASSETT, D. S. and BULLMORE, E. T. (2009). Human brain networks in health and disease. *Current Opinion in Neurology* **22** 340–347.
- BELILOVSKY, E., VAROQUAUX, G. and BLASCHKO, M. B. (2016). Testing for differences in Gaussian graphical models: Applications to brain connectivity. In *Advances in Neural Information Processing Systems* (D. D. Lee, M. Sugiyama, U. V. Luxberg, I. Guyon and R. Garnett, eds.) **29** 595–603. Curran Associates, Red Hook, NY.
- BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. [MR3037163 https://doi.org/10.3150/11-BEJ410](https://doi.org/10.3150/11-BEJ410)
- BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. [MR2533469 https://doi.org/10.1214/08-AOS620](https://doi.org/10.1214/08-AOS620)
- BUSCH, S., ACAR, A., MAGNUSSON, Y., GREGERSSON, P., RYDÉN, L. and LANDBERG, G. (2015). TGF-beta receptor type-2 expression in cancer-associated fibroblasts regulates breast cancer cell growth and survival and is a prognostic marker in pre-menopausal breast cancer. *Oncogene* **34** 27–38. <https://doi.org/10.1038/onc.2013.527>
- CAREY, L. A., PEROU, C. M., LIVASY, C. A., DRESSLER, L. G., COWAN, D., CONWAY, K., KARACA, G., TROESTER, M. A., TSE, C. K. et al. (2006). Race, breast cancer subtypes, and survival in the Carolina breast cancer study. *JAMA* **295** 2492–2502.
- CARVALHO, I., MILANEZI, F., MARTINS, A., REIS, R. M. and SCHMITT, F. (2005). Overexpression of platelet-derived growth factor receptor alpha in breast cancer is associated with tumour progression. *Breast Cancer Res.* **7** R788–R795.
- DANAHER, P., WANG, P. and WITTEN, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 373–397. [MR3164871 https://doi.org/10.1111/rssb.12033](https://doi.org/10.1111/rssb.12033)
- FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. [MR1946581 https://doi.org/10.1198/016214501753382273](https://doi.org/10.1198/016214501753382273)
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

Key words and phrases. Differential connectivity, biological networks, significance test, lasso, high-dimensional data.

- GILL, R., DATTA, S. and DATTA, S. (2014). dna: An R package for differential network analysis. *Bioinformatics* **10** 233–234.
- GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1–15. [MR2804206 https://doi.org/10.1093/biomet/asq060](https://doi.org/10.1093/biomet/asq060)
- HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6** 65–70. [MR0538597](https://doi.org/10.2307/2346977)
- IDEKER, T. and KROGAN, N. J. (2012). Differential network biology. *Mol. Syst. Biol.* **8** 565.
- JANES, K. A. (2011). RUNX1 and its understudied role in breast cancer. *Cell Cycle* **10** 3461–3465.
- JANKOVÁ, J. and VAN DE GEER, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Stat.* **9** 1205–1229. [MR3354336 https://doi.org/10.1214/15-EJS1031](https://doi.org/10.1214/15-EJS1031)
- JANKOVÁ, J. and VAN DE GEER, S. (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *TEST* **26** 143–162. [MR3613609 https://doi.org/10.1007/s11749-016-0503-5](https://doi.org/10.1007/s11749-016-0503-5)
- JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. [MR3277152](https://doi.org/10.26434/chemrxiv-2014-05)
- JEZIEBSKA, A. and MOTYL, T. (2009). Matrix metalloproteinase-2 involvement in breast cancer progression: A mini-review. *Med. Sci. Monit.* **15** RA32–RA40.
- JÖNSSON, G., STAAF, J., VALLON-CHRISTERSSON, J., RINGNÉR, M., HOLM, K., HEGARDT, C., GUNNARSSON, H., FAGERHOLM, R., STRAND, C. et al. (2010). Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Res.* **12** 1–14.
- KITANO, H. (2004). Biological robustness. *Nat. Rev. Genet.* **5** 826–837.
- KRUMSIEK, J., SUHRE, K., ILLIG, T., ADAMSKI, J. and THEIS, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.* **5** 21. <https://doi.org/10.1186/1752-0509-5-21>
- LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. [MR3485948 https://doi.org/10.1214/15-AOS1371](https://doi.org/10.1214/15-AOS1371)
- LEEB, H. and PÖTSCHER, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* **24** 338–376. [MR2422862 https://doi.org/10.1017/S0266466608080158](https://doi.org/10.1017/S0266466608080158)
- 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](https://doi.org/10.1111/j.1541-0420.2007.00799.x)
- MA, X., BEEGLY-FADIEL, A., LU, W., SHI, J., XIANG, Y. B., CAI, Q., SHEN, H., SHEN, C. Y., REN, Z. et al. (2012). Pathway analyses identify TGFBR2 as potential breast cancer susceptibility gene: Results from a consortium study among asians. *Cancer Epidemiol. Biomark. Prev.* **21** 1176–1187.
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. [MR2278363 https://doi.org/10.1214/009053606000000281](https://doi.org/10.1214/009053606000000281)
- MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). *p*-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681. [MR2750584 https://doi.org/10.1198/jasa.2009.tm08647](https://doi.org/10.1198/jasa.2009.tm08647)
- NEWMAN, M. E. J. (2003). The structure and function of complex networks. *SIAM Rev.* **45** 167–256. [MR2010377 https://doi.org/10.1137/S003614450342480](https://doi.org/10.1137/S003614450342480)
- NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45** 158–195. [MR3611489 https://doi.org/10.1214/16-AOS1448](https://doi.org/10.1214/16-AOS1448)
- PELLEGRINI, R., MARTIGNONE, S., TAGLIABUE, E., BELOTTI, D., BUFALINO, R., CASCINELLI, N., MÉNARD, S. and COLNAGHI, M. I. (1995). Prognostic significance of laminin production in relation with its receptor expression in human breast carcinomas. *Cancer Res.* **35** 195–199.
- PEROU, C. M., SØRLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H. et al. (2000). Molecular portraits of human breast tumours. *Nature* **406** 747–752.
- PETERSON, C., STINGO, F. C. and VANNUCCI, M. (2015). Bayesian inference of multiple Gaussian graphical models. *J. Amer. Statist. Assoc.* **110** 159–174. [MR3338494 https://doi.org/10.1080/01621459.2014.896806](https://doi.org/10.1080/01621459.2014.896806)
- REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. [MR3346695 https://doi.org/10.1214/14-AOS1286](https://doi.org/10.1214/14-AOS1286)
- SAEGUSA, T. and SHOJAIE, A. (2016). Joint estimation of precision matrices in heterogeneous populations. *Electron. J. Stat.* **10** 1341–1392. [MR3507368 https://doi.org/10.1214/16-EJS1137](https://doi.org/10.1214/16-EJS1137)
- SAWYER, C., STURGE, J., BENNETT, D. C., O'HARE, M. J., ALLEN, W. E., BAIN, J., JONES, G. E. and VANHAESEBROECK, B. (2003). Regulation of breast cancer cell chemotaxis by the phosphoinositide 3-kinase p110delta. *Cancer Res.* **63** 1667–1675.
- STÄDLER, N. and MUKHERJEE, S. (2017). Two-sample testing in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 225–246. [MR3597971 https://doi.org/10.1111/rssb.12173](https://doi.org/10.1111/rssb.12173)

- TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. and TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.* **111** 600–620. [MR3538689](#) <https://doi.org/10.1080/01621459.2015.1108848>
- VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. [MR3224285](#) <https://doi.org/10.1214/14-AOS1221>
- WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. [MR2543689](#) <https://doi.org/10.1214/08-AOS646>
- XIA, Y., CAI, T. and CAI, T. T. (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika* **102** 247–266. [MR3371002](#) <https://doi.org/10.1093/biomet/asu074>
- XIA, Y. and LI, L. (2017). Hypothesis testing of matrix graph model with application to brain connectivity analysis. *Biometrics* **73** 780–791. [MR3713112](#) <https://doi.org/10.1111/biom.12633>
- ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. [MR2604701](#) <https://doi.org/10.1214/09-AOS729>
- ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. [MR3153940](#) <https://doi.org/10.1111/rssb.12026>
- ZHAO, S. D., CAI, T. T. and LI, H. (2014). Direct estimation of differential networks. *Biometrika* **101** 253–268. [MR3215346](#) <https://doi.org/10.1093/biomet/asu009>
- ZHAO, S. and SHOJAIE, A. (2016). A significance test for graph-constrained estimation. *Biometrics* **72** 484–493. [MR3515775](#) <https://doi.org/10.1111/biom.12418>
- ZHAO, S. and SHOJAIE, A. (2022). Supplement to “Network differential connectivity analysis.” <https://doi.org/10.1214/21-AOAS1581SUPPA>, <https://doi.org/10.1214/21-AOAS1581SUPPB>
- ZHAO, S., WITTEN, D. and SHOJAIE, A. (2021). In defense of the indefensible: A very naïve approach to high-dimensional inference. *Statist. Sci.* **36** 562–577. [MR4323053](#) <https://doi.org/10.1214/20-sts815>
- ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. [MR2274449](#)

ESTIMATING HETEROGENEOUS GENE REGULATORY NETWORKS FROM ZERO-INFLATED SINGLE-CELL EXPRESSION DATA

BY QIUYU WU^a AND XIANGYU LUO^b

Institute of Statistics and Big Data, Renmin University of China, ^aw.qy@ruc.edu.cn, ^bxiangyuluo@ruc.edu.cn

Inferring gene regulatory networks can elucidate how genes work cooperatively. The gene-gene collaboration information is often learned by Gaussian graphical models (GGM) that aim to identify whether the expression levels of any pair of genes are dependent, given other genes' expression values. One basic assumption that guarantees the validity of GGM is data normality, and this often holds for *bulk-level* expression data which aggregate biological signals from a collection of cells. However, fine-grained *cell-level* expression profiles collected in single-cell RNA-sequencing (scRNA-seq) reveal non-normality features—cellular heterogeneity and zero inflation. We propose a Bayesian latent mixture GGM to jointly estimate multiple gene regulatory networks accounting for the zero inflation and unknown heterogeneity of single-cell expression data. The proposed approach outperforms competing methods on synthetic data in terms of network structure and precision matrix estimation accuracy and provides biological insights when applied to two real-world scRNA-seq datasets. An R package implementing the proposed model is available on GitHub <https://github.com/WgitU/BLGGM>.

REFERENCES

- AIBAR, S., GONZÁLEZ-BLAS, C. B., MOERMAN, T., IMRICHÓVA, H., HULSELMANS, G., RAMBOW, F., MARINE, J.-C., GEURTS, P., AERTS, J. et al. (2017). SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **14** 1083–1086.
- AMEMIYA, T. (1984). Tobit models: A survey. *J. Econometrics* **24** 3–61. [MR0739428 https://doi.org/10.1016/0304-4076\(84\)90074-5](https://doi.org/10.1016/0304-4076(84)90074-5)
- 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](https://doi.org/10.1007/978-0-387-45528-0)
- CHATTERJEE, S., KAPOOR, A., AKIYAMA, J. A., AUER, D. R., LEE, D., GABRIEL, S., BERRIOS, C., PENNACCHIO, L. A. and CHAKRAVARTI, A. (2016). Enhancer variants synergistically drive dysfunction of a gene regulatory network in Hirschsprung disease. *Cell* **167** 355–368.
- CHEN, M. and ZHOU, X. (2017). Controlling for confounding effects in single cell RNA sequencing studies using both control and target genes. *Sci. Rep.* **7** 1–14.
- CHOI, M. J., TAN, V. Y. F., ANANDKUMAR, A. and WILLSKY, A. S. (2011). Learning latent tree graphical models. *J. Mach. Learn. Res.* **12** 1771–1812. [MR2813153 https://doi.org/10.1111/rssb.12033](https://doi.org/10.1111/rssb.12033)
- CSARDI, G., NEPUSZ, T. et al. (2006). The igraph software package for complex network research. *InterJournal, Complex Syst.* **1695** 1–9.
- DANAHER, P., WANG, P. and WITTEN, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 373–397. [MR3164871 https://doi.org/10.1111/rssb.12033](https://doi.org/10.1111/rssb.12033)
- DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175. [MR3931974 https://doi.org/10.1111/rssb.12033](https://doi.org/10.1111/rssb.12033)
- DOBRA, A., LENKOSKI, A. and RODRIGUEZ, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *J. Amer. Statist. Assoc.* **106** 1418–1433. [MR2896846 https://doi.org/10.1198/jasa.2011.tm10465](https://doi.org/10.1198/jasa.2011.tm10465)
- EDGAR, R., DOMRACHEV, M. and LASH, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30** 207–210.
- 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](https://doi.org/10.1198/016214502760047131)

Key words and phrases. Bayesian analysis, Gaussian graphical models, heterogeneity, nonignorable dropout, spike-slab prior.

- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.
- GAJOVIĆ, S., MITREČIĆ, D., AUGUSTINČIĆ, L., IACONCIG, A. and MURO, A. F. (2006). Unexpected rescue of alpha-synuclein and multimerin1 deletion in C57BL/6JOLA^{Hsd} mice by beta-adducin knockout. *Transgenic Res.* **15** 255–259.
- GALLOPIN, M., RAU, A. and JAFFRÉZIC, F. (2013). A hierarchical Poisson log-normal model for network inference from RNA sequencing data. *PLoS ONE* **8** e77503. <https://doi.org/10.1371/journal.pone.0077503>
- GAN, L., YANG, X., NARISSETTY, N. and LIANG, F. (2019). Bayesian joint estimation of multiple graphical models. In *Advances in Neural Information Processing Systems* 9799–9809.
- GAO, C., ZHU, Y., SHEN, X. and PAN, W. (2016). Estimation of multiple networks in Gaussian mixture models. *Electron. J. Stat.* **10** 1133–1154. [MR3499523 https://doi.org/10.1214/16-EJS1135](https://doi.org/10.1214/16-EJS1135)
- GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. [MR3235677](https://doi.org/10.1214/13-BA889)
- GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.
- GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1–15. [MR2804206 https://doi.org/10.1093/biomet/asq060](https://doi.org/10.1093/biomet/asq060)
- HAO, B., SUN, W. W., LIU, Y. and CHENG, G. (2018). Simultaneous clustering and estimation of heterogeneous graphical models. *J. Mach. Learn. Res.* **18** Paper No. 217, 58 pp. [MR3827105](https://doi.org/10.1162/jmlr-2018-18-021)
- HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.
- KIM, S. (2015). ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Commun. Stat. Appl. Methods* **22** 665.
- KOHLSCHEEN, S., WINTERLE, S., SCHWARZER, A., KAMP, C., BRUGMAN, M. H., BREUER, D. C., BÜSCHE, G., BAUM, C. and MODLICH, U. (2015). Inhibition of Thrombopoietin/Mpl signaling in adult hematopoiesis identifies new candidates for hematopoietic stem cell maintenance. *PLoS ONE* **10** e0131866. <https://doi.org/10.1371/journal.pone.0131866>
- LI, Z., MCCORMICK, T. and CLARK, S. (2019). Bayesian joint spike-and-slab graphical lasso. In *International Conference on Machine Learning* 3877–3885.
- LIN, Z., WANG, T., YANG, C. and ZHAO, H. (2017). On joint estimation of Gaussian graphical models for spatial and temporal data. *Biometrics* **73** 769–779. [MR3713111 https://doi.org/10.1111/biom.12650](https://doi.org/10.1111/biom.12650)
- LIN, Z., ZAMANIGHOMI, M., DALEY, T., MA, S. and WONG, W. H. (2020). Model-based approach to the joint analysis of single-cell data on chromatin accessibility and gene expression. *Statist. Sci.* **35** 2–13. [MR4071354 https://doi.org/10.1214/19-STS714](https://doi.org/10.1214/19-STS714)
- LUO, X. and WEI, Y. (2018). Nonparametric Bayesian learning of heterogeneous dynamic transcription factor networks. *Ann. Appl. Stat.* **12** 1749–1772. [MR3852696 https://doi.org/10.1214/17-AOAS1129](https://doi.org/10.1214/17-AOAS1129)
- MA, J. and MICHAILIDIS, G. (2016). Joint structural estimation of multiple graphical models. *J. Mach. Learn. Res.* **17** Paper No. 166, 48 pp. [MR3555057](https://doi.org/10.1162/jmlr-2016-17-016)
- MCDAVID, A., GOTTARDO, R., SIMON, N. and DRTON, M. (2019). Graphical models for zero-inflated single cell gene expression. *Ann. Appl. Stat.* **13** 848–873. [MR3963555 https://doi.org/10.1214/18-AOAS1213](https://doi.org/10.1214/18-AOAS1213)
- MENG, Z., ERIKSSON, B. and HERO, A. (2014). Learning latent variable Gaussian graphical models. In *International Conference on Machine Learning* 1269–1277. PMLR.
- MENON, M., MOHAMMADI, S., DAVILA-VELDERRAIN, J., GOODS, B. A., CADWELL, T. D., XING, Y., STEMMER-RACHAMIMOV, A., SHALEK, A. K., LOVE, J. C. et al. (2019). Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration. *Nat. Commun.* **10** 1–9.
- METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.
- MIAO, W., DING, P. and GENG, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *J. Amer. Statist. Assoc.* **111** 1673–1683. [MR3601726 https://doi.org/10.1080/01621459.2015.1105808](https://doi.org/10.1080/01621459.2015.1105808)
- MOHAMMADI, A. and WIT, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Anal.* **10** 109–138. [MR3420899 https://doi.org/10.1214/14-BA889](https://doi.org/10.1214/14-BA889)
- MOHAMMADI, R. and WIT, E. C. (2019). BDgraph: An R package for Bayesian structure learning in graphical models. *J. Stat. Softw.* **89** 1–30.
- MOHAMMADI, S., RAVINDRA, V., GLEICH, D. F. and GRAMA, A. (2018). A geometric approach to characterize the functional identity of single cells. *Nat. Commun.* **9** 1–10.
- MORITA, Y., EMA, H. and NAKAUCHI, H. (2010). Heterogeneity and hierarchy within the most primitive hematopoietic stem cell compartment. *J. Exp. Med.* **207** 1173–1182.
- NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 113–162. CRC Press, Boca Raton, FL. [MR2858447](https://doi.org/10.1214/11-HMC113)

- NESTOROWA, S., HAMEY, F. K., PIJUAN SALA, B., DIAMANTI, E., SHEPHERD, M., LAURENTI, E., WILSON, N. K., KENT, D. G. and GÖTTGENS, B. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood, J. Amer. Soc. Hematol.* **128** e20–e31.
- NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.
- NTRANOS, V., YI, L., MELSTED, P. and PACTER, L. (2019). A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat. Methods* **16** 163–166. <https://doi.org/10.1038/s41592-018-0303-9>
- PAN, W. and SHEN, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* **8** 1145–1164.
- PETERSON, C., STINGO, F. C. and VANNUCCI, M. (2015). Bayesian inference of multiple Gaussian graphical models. *J. Amer. Statist. Assoc.* **110** 159–174. [MR3338494 https://doi.org/10.1080/01621459.2014.896806](https://doi.org/10.1080/01621459.2014.896806)
- PIERSON, E. and YAU, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16** 1–10.
- PRATAPA, A., JALIHAI, A. P., LAW, J. N., BHARADWAJ, A. and MURALI, T. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17** 147–154.
- QIU, X., RAHIMZAMANI, A., WANG, L., MAO, Q., DURHAM, T., MCFALINE-FIGUEROA, J. L., SAUNDERS, L., TRAPNELL, C. and KANNAN, S. (2018). Towards inferring causal gene regulatory networks from single cell expression measurements. *BioRxiv* 426981.
- REN, M., ZHANG, S., ZHANG, Q. and MA, S. (2021a). Gaussian graphical model-based heterogeneity analysis via penalized fusion. *Biometrics*. <https://doi.org/10.1111/biom.13426>
- REN, M., ZHANG, S., ZHANG, Q. and MA, S. (2021b). HeteroGGM: An R package for Gaussian graphical model-based heterogeneity analysis. *Bioinformatics* **37** 3073–3074.
- RISSE, D., PERRAUDEAU, F., GRIBKOVA, S., DUDOIT, S. and VERT, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9** 1–17.
- RODRÍGUEZ, A., LENKOSKI, A. and DOBRA, A. (2011). Sparse covariance estimation in heterogeneous samples. *Electron. J. Stat.* **5** 981–1014. [MR2836767 https://doi.org/10.1214/11-EJS634](https://doi.org/10.1214/11-EJS634)
- ROZENBLATT-ROSEN, O., STUBBINGTON, M. J., REGEV, A. and TEICHMANN, S. A. (2017). The Human Cell Atlas: From vision to reality. *Nat. News* **550** 451.
- SAEGUSA, T. and SHOJAIE, A. (2016). Joint estimation of precision matrices in heterogeneous populations. *Electron. J. Stat.* **10** 1341–1392. [MR3507368 https://doi.org/10.1214/16-EJS1137](https://doi.org/10.1214/16-EJS1137)
- SONG, F., CHAN, G. M. A. and WEI, Y. (2020). Flexible experimental designs for valid single-cell RNA-sequencing experiments allowing batch effects correction. *Nat. Commun.* **11** 1–15.
- TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. [MR0898357](https://doi.org/10.1080/01621459.1987.1571745)
- VAN GENDEREN, M. M., BIJVELD, M. M., CLAASSEN, Y. B., FLORIJN, R. J., PEARRING, J. N., MEIRE, F. M., MCCALL, M. A., RIEMSLAG, F. C., GREGG, R. G. et al. (2009). Mutations in TRPM1 are a common cause of complete congenital stationary night blindness. *Am. J. Hum. Genet.* **85** 730–736.
- VIETH, B., PAREKH, S., ZIEGENHAIN, C., ENARD, W. and HELLMANN, I. (2019). A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.* **10** 1–11.
- WANG, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Anal.* **7** 867–886. [MR3000017 https://doi.org/10.1214/12-BA729](https://doi.org/10.1214/12-BA729)
- WANG, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Anal.* **10** 351–377. [MR3420886 https://doi.org/10.1214/14-BA916](https://doi.org/10.1214/14-BA916)
- WANG, H. and LI, S. Z. (2012). Efficient Gaussian graphical model determination under G -Wishart prior distributions. *Electron. J. Stat.* **6** 168–198. [MR2879676 https://doi.org/10.1214/12-EJS669](https://doi.org/10.1214/12-EJS669)
- WU, Q. and LUO, X. (2022). Supplement to “Estimating heterogeneous gene regulatory networks from zero-inflated single-cell expression data.” <https://doi.org/10.1214/21-AOAS1582SUPPA>, <https://doi.org/10.1214/21-AOAS1582SUPPB>
- YANG, H.-J., RATNAPRIYA, R., COGLIATI, T., KIM, J.-W. and SWAROOP, A. (2015). Vision from next generation sequencing: Multi-dimensional genome-wide analysis for producing gene regulatory networks underlying retinal development, aging and disease. *Prog. Retin. Eye Res.* **46** 1–30.
- YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. [MR2367824 https://doi.org/10.1093/biomet/asm018](https://doi.org/10.1093/biomet/asm018)
- ZHANG, H., XU, J., JIANG, N., HU, X. and LUO, Z. (2015). PLNseq: A multivariate Poisson lognormal distribution for high-throughput matched RNA-sequencing read count data. *Stat. Med.* **34** 1577–1589. [MR3334677 https://doi.org/10.1002/sim.6449](https://doi.org/10.1002/sim.6449)

ACCOUNTING FOR SURVEY DESIGN IN BAYESIAN DISAGGREGATION OF SURVEY-BASED AREAL ESTIMATES OF PROPORTIONS: AN APPLICATION TO THE AMERICAN COMMUNITY SURVEY

BY MARCO H. BENEDETTI^{1,a}, VERONICA J. BERROCAL^{2,b} AND RODERICK J. LITTLE^{3,c}

¹Nationwide Children's Hospital, Center for Injury Research and Policy, ^amarco.benedetti@nationwidechildrens.org

²Department of Statistics, School of Information and Computer Sciences, University of California, Irvine, ^bvberroca@uci.edu

³Department of Biostatistics, School of Public Health, University of Michigan, ^crlittle@umich.edu

Understanding the effects of social determinants of health on health outcomes requires data on characteristics of the neighborhoods in which subjects live. However, estimates of these characteristics are often aggregated over space and time in a fashion that diminishes their utility. Take, for example, estimates from the American Community Survey (ACS), a multiyear nationwide survey administered by the U.S. Census Bureau: estimates for small municipal areas are aggregated over 5-year periods, whereas 1-year estimates are only available for municipal areas with populations >65,000. Researchers may wish to use ACS estimates in studies of population health to characterize neighborhood-level exposures. However, 5-year estimates may not properly characterize temporal changes or align temporally with other data in the study, while the coarse spatial resolution of the 1-year estimates diminishes their utility in characterizing neighborhood exposure. To circumvent this issue, in this paper we propose a modeling framework to disaggregate estimates of proportions derived from sampling surveys, which explicitly accounts for the survey design effect. We illustrate the utility of our model by applying it to the ACS data, generating estimates of poverty for the state of Michigan at fine spatiotemporal resolution.

REFERENCES

- ABRAMS, E. M. and SZEFLER, S. J. (2020). COVID-19 and the impact of social determinants of health. *Lancet Respir. Med.* **8** 659–661. [https://doi.org/10.1016/S2213-2600\(20\)30234-4](https://doi.org/10.1016/S2213-2600(20)30234-4)
- AGUILAR, L. (2015). Detroit's Cass Corridor makes way for new era. The Detroit News, published April 2015.
- ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. [MR1224394](https://doi.org/10.1080/01621459.1993.10477394)
- BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, FL.
- BENEDETTI, M. H., BERROCAL, V. J. and LITTLE, R. J. (2022). Supplement to “Accounting for survey design in Bayesian disaggregation of survey-based areal estimates of proportions: an application to the American Community Survey.” <https://doi.org/10.1214/21-AOAS1585SUPPA>, <https://doi.org/10.1214/21-AOAS1585SUPPB>
- BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2010). A bivariate space–time downscaler under space and time misalignment. *Ann. Appl. Stat.* **4** 1942–1975. [MR2829942](https://doi.org/10.1214/10-AOAS351) <https://doi.org/10.1214/10-AOAS351>
- BRADLEY, J. R., HOLAN, S. H. and WIKLE, C. K. (2016). Multivariate spatio-temporal survey fusion with application to the American Community Survey and Local Area Unemployment Statistics. *Stat* **5** 224–233. [MR3564657](https://doi.org/10.1002/sta4.120) <https://doi.org/10.1002/sta4.120>
- BRADLEY, J. R., WIKLE, C. K. and HOLAN, S. H. (2015). Spatio-temporal change of support with application to American Community Survey multi-year period estimates. *Stat* **4** 255–270. [MR3414659](https://doi.org/10.1002/sta4.94) <https://doi.org/10.1002/sta4.94>
- BRADLEY, J. R., WIKLE, C. K. and HOLAN, S. H. (2016). Bayesian spatial change of support for count-valued survey data with application to the American community survey. *J. Amer. Statist. Assoc.* **111** 472–487. [MR3538680](https://doi.org/10.1080/01621459.2015.1117471) <https://doi.org/10.1080/01621459.2015.1117471>

Key words and phrases. Spatiotemporal change of support problem, Bayesian hierarchical model, multi-resolution approximation, latent spatiotemporal process, American Community Survey, survey-based estimates.

- BRAVERMAN, P., EGERTER, S. and WILLIAMS, D. R. (2011). The social determinants of health: Coming of age. *Annual Reviews of Public Health* **32** 381–398.
- CHEN, C., WAKEFIELD, J. and LUMLEY, T. (2014). The use of sampling weights in Bayesian hierarchical models for small area estimation. *Spat. Spatio-Temporal. Epidemiol.* **11** 33–43.
- DAWID, A. P. (1982). The well-calibrated Bayesian. *J. Amer. Statist. Assoc.* **77** 605–613. [MR0675887](#)
- DIGGLE, P. J., TAWN, J. A. and MOYEED, R. A. (1998). Model-based geostatistics. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **47** 299–350. [MR1626544](#) <https://doi.org/10.1111/1467-9876.00113>
- DURANTE, D. (2019). Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika* **106** 765–779. [MR4031198](#) <https://doi.org/10.1093/biomet/asz034>
- ENTEZARI, R., BROWN, P. E. and ROSENTHAL, J. S. (2019). Bayesian spatial analysis of hardwood tree counts via MCMC. *Environmetrics* **31** e2608.
- FAY, R. E. III and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277. [MR0548019](#)
- FINLEY, A., BANERJEE, S. and CARLIN, B. P. (2007). spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *J. Stat. Softw.* **19** 1–24.
- GELFAND, A. E., BANERJEE, S. and GAMERMAN, D. (2005). Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics* **16** 465–479. [MR2147537](#) <https://doi.org/10.1002/env.715>
- GELFAND, A. E., ZHU, L. and CARLIN, B. P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics* **2** 31–45.
- GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, 4 (Peñíscola, 1991) J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 169–193. Oxford Univ. Press, New York. [MR1380276](#)
- GHITZA, Y. and GELMAN, A. (2013). Deep interaction with MRP: Election turnout and voting patterns among small electoral subgroups. *Amer. J. Polit. Sci.* **57** 762–776.
- GILANI, O., BERROCAL, V. J. and BATTERMAN, S. A. (2019). Nonstationary spatiotemporal Bayesian data fusion for pollutants in the near-road environment. *Environmetrics* **30** e2581. [MR4021438](#) <https://doi.org/10.1002/env.2581>
- GOTWAY, C. A. and YOUNG, L. J. (2002). Combining incompatible spatial data. *J. Amer. Statist. Assoc.* **97** 632–648. [MR1951636](#) <https://doi.org/10.1198/016214502760047140>
- HEATON, M. J., DATTA, A., FINLEY, A. O., FURRER, R., GUINNESS, J., GUHANIYOGI, R., GERBER, F., GRAMACY, R. B., HAMMERLING, D. et al. (2019). A case study competition among methods for analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* **24** 398–425.
- IACHAN, R., PIERANNUNZI, C., HEALEY, K., GREENLUND, K. and TOWN, M. (2016). National weighting of data from the behavioral risk factor surveillance system (BRFSS). *BMC Med. Res. Methodol.* **16** 1–12.
- KATZFUSS, M. (2017). A multi-resolution approximation for massive spatial datasets. *J. Amer. Statist. Assoc.* **112** 201–214. [MR3646566](#) <https://doi.org/10.1080/01621459.2015.1123632>
- KISH, L. (1965). *Survey Sampling*. Wiley, New York.
- KISH, L. (1995). Methods for design effects. *J. Off. Stat.* **11** 55–77.
- KORN, E. L. and GRAUBARD, B. I. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Surv. Methodol.* **24** 193–201.
- LI, Z., HSIAO, Y., GODWIN, J., MARTIN, B. D., WAKEFIELD, J. and CLARK, S. J. (2019). Changes in the spatial distribution of the under-five mortality rate: Small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. *PLoS ONE* **14** 1–17.
- LITTLE, R. J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *Amer. Statist.* **60** 213–223. [MR2246754](#) <https://doi.org/10.1198/000313006X117837>
- MARMOT, M., ALLEN, J., BELL, R., BLOOMER, E., GOLDBLATT, P., CONSORTIUM FOR THE EUROPEAN REVIEW OF SOCIAL DETERMINANTS OF HEALTH AND THE HEALTH DIVIDE (2012). WHO European review of social determinants of health and the health divide. *Lancet* **380** 1011–1029.
- MERCER, L., WAKEFIELD, J., CHEN, C. and LUMLEY, T. (2014). A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spat. Stat.* **8** 69–85. [MR3326822](#) <https://doi.org/10.1016/j.spasta.2013.12.001>
- MITCHELL, M. W., GENTON, M. G. and GUMPERTZ, M. L. (2005). Testing for separability of space–time covariances. *Environmetrics* **16** 819–831. [MR2216653](#) <https://doi.org/10.1002/env.737>
- MOEHLMAN, L. and ROBINS-SOMERVILLE, M. (2016). The new Detroit: How gentrification has changed Detroit’s economic landscape. Michigan Daily, published September 2016.
- PEREIRA, L. N. and COELHO, P. S. (2010). Small area estimation of mean price of habitation transaction using time-series and cross-sectional area-level models. *J. Appl. Stat.* **37** 651–666. [MR2756658](#) <https://doi.org/10.1080/02664760902810821>
- PFEFFERMANN, D. (2013). New important developments in small area estimation. *Statist. Sci.* **28** 40–68. [MR3075338](#) <https://doi.org/10.1214/12-STS395>

- 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>
- PORTER, A. T., HOLAN, S. H., WIKLE, C. K. and CRESSIE, N. (2014). Spatial Fay–Herriot models for small area estimation with functional covariates. *Spat. Stat.* **10** 27–42. [MR3280088](#) <https://doi.org/10.1016/j.spasta.2014.07.001>
- PRATESI, M. and SALVATI, N. (2008). Small area estimation: The EBLUP estimator based on spatially correlated random area effects. *Stat. Methods Appl.* **17** 113–141. [MR2416883](#) <https://doi.org/10.1007/s10260-007-0061-9>
- ROBERTS, G. O., GELMAN, A. and GILKS, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7** 110–120. [MR1428751](#) <https://doi.org/10.1214/aoap/1034625254>
- ROLLSTON, R. and GALEA, S. (2020). COVID-19 and the social determinants of health. *Am. J. Health Promot.* **34** 687–689. <https://doi.org/10.1177/0890117120930536b>
- SAVITSKY, T. D. (2016). Bayesian nonparametric multiresolution estimation for the American Community Survey. *Ann. Appl. Stat.* **10** 2157–2181. [MR3592052](#) <https://doi.org/10.1214/16-AOAS968>
- 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>
- SIMPSON, M., HOLAN, S. H., WIKLE, C. K. and BRADLEY, J. R. (2019). Interpolating distributions for populations in nested geographies using public-use data with application to the American Community Survey. Preprint. Available at [arXiv:1802.02626](https://arxiv.org/abs/1802.02626).
- SINGH, B., SHUKLA, G. and KUNDU, D. (2005). Spatio-temporal models in small-area estimation. *Surv. Methodol.* **31** 183–195.
- SINGU, S., ACHARYA, A., CHALLAGUNDLA, K. and BYAREDDY, S. B. (2020). Impact of social determinants of health on the emerging COVID-19 pandemic in the United States. *Frontiers in Public Health* **8** 406.
- SØRBYE, S. H. and RUE, H. (2011). Simultaneous credible bands for latent Gaussian models. *Scand. J. Stat.* **38** 712–725. [MR2859746](#) <https://doi.org/10.1111/j.1467-9469.2011.00741.x>
- U.S. CENSUS BUREAU (2008). A Compass for Understanding and Using American Community Survey Data: What General Data Users Need to Know. U.S. Government Printing Office, Washington, DC.
- U.S. CENSUS BUREAU (2014). American Community Survey Design and Methodology. U.S. Government Printing Office, Washington, DC.

BAYESIAN MULTIVARIATE SPARSE FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS WITH APPLICATION TO LONGITUDINAL MICROBIOME MULTIOMICS DATA

BY LINGJING JIANG^{1,a}, CHRIS ELROD^{2,c}, JANE J. KIM^{3,d}, AUSTIN D. SWAFFORD^{4,e},
ROB KNIGHT^{5,f} AND WESLEY K. THOMPSON^{1,b}

¹Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego,

^alij014@health.ucsd.edu, ^bwkthompson@health.ucsd.edu

²Julia Computing, ^cchris.elrod@juliacomputing.com

³Department of Pediatrics, University of California San Diego, ^djanekim@health.ucsd.edu

⁴Center for Microbiome Innovation, University of California San Diego, ^eadswafford@eng.ucsd.edu

⁵Department of Pediatrics, Center for Microbiome Innovation, Department of Computer Science and Engineering, Department of Bioengineering, University of California San Diego, ^frobknight@health.ucsd.edu

Microbiome researchers often need to model the temporal dynamics of multiple complex, nonlinear outcome trajectories simultaneously. This motivates our development of *multivariate Sparse Functional Principal Components Analysis* (mSFPCA), extending existing SFPCA methods to simultaneously characterize multiple temporal trajectories and their interrelationships. As with existing SFPCA methods, the mSFPCA algorithm characterizes each trajectory as a smooth mean plus a weighted combination of the smooth major modes of variation about the mean, where the weights are given by the component scores for each subject. Unlike existing SFPCA methods, the mSFPCA algorithm allows estimation of multiple trajectories simultaneously, such that the component scores, which are constrained to be independent within a particular outcome for identifiability, may be arbitrarily correlated with component scores for other outcomes. A Cholesky decomposition is used to estimate the component score covariance matrix efficiently and guarantee positive semidefiniteness given these constraints. Mutual information is used to assess the strength of marginal and conditional temporal associations across outcome trajectories. Importantly, we implement mSFPCA as a Bayesian algorithm using R and stan, enabling easy use of packages such as PSIS-LOO for model selection and graphical posterior predictive checks to assess the validity of mSFPCA models. Although we focus on application of mSFPCA to microbiome data in this paper, the mSFPCA model is of general utility and can be used in a wide range of real-world applications.

REFERENCES

- INTEGRATIVE HMP (IHMP) RESEARCH NETWORK CONSORTIUM (2014). The Integrative Human Microbiome Project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* **16** 276–289. <https://doi.org/10.1016/j.chom.2014.08.014>
- AHMED, N. A. and GOKHALE, D. V. (1989). Entropy expressions and their estimators for multivariate distributions. *IEEE Trans. Inf. Theory* **35** 688–692. [MR1022093 https://doi.org/10.1109/18.30996](https://doi.org/10.1109/18.30996)
- ARELLANO-VALLE, R. B., CONTRERAS-REYES, J. E. and GENTON, M. G. (2013). Shannon entropy and mutual information for multivariate skew-elliptical distributions. *Scand. J. Stat.* **40** 42–62. [MR3024031 https://doi.org/10.1111/j.1467-9469.2011.00774.x](https://doi.org/10.1111/j.1467-9469.2011.00774.x)
- BALLEN, K., AHN, K. W., CHEN, M., ABDEL-AZIM, H., AHMED, I., ALJURF, M., ANTIN, J., BHATT, A. S., BOECKH, M. et al. (2016). Infection rates among acute leukemia patients receiving alternative donor hematopoietic cell transplantation. *Biol. Blood Marrow Transplant.* **22** 1636–1645.
- BARNARD, J., MCCULLOCH, R. and MENG, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist. Sinica* **10** 1281–1311. [MR1804544](https://doi.org/10.1007/BF02429544)

- BODEIN, A., CHAPLEUR, O., DROIT, A. and LÊ CAO, K.-A. (2019). A generic multivariate framework for the integration of microbiome longitudinal studies with other data types. *Front. Genet.* **10** 963.
- BOUSLIMANI, A., DA SILVA, R., KOSCIOLEK, T., JANSSEN, S., CALLEWAERT, C., AMIR, A., DORRESTEIN, K., MELNIK, A. V., ZARAMELA, L. S. et al. (2019). The impact of skin care products on skin chemistry and microbiome dynamics. *BMC Biol.* **17** 1–20.
- 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** 1–32.
- CHIOU, J.-M., CHEN, Y.-T. and YANG, Y.-F. (2014). Multivariate functional principal component analysis: A normalization approach. *Statist. Sinica* **24** 1571–1596. [MR3308652](#)
- COVER, T. M. (1999). *Elements of Information Theory*. Wiley, New York.
- DI, C., CRAINICEANU, C. M. and JANK, W. S. (2014). Multilevel sparse functional principal component analysis. *Stat* **3** 126–143. [MR4027332](#) <https://doi.org/10.1002/sta4.50>
- DI, C.-Z., CRAINICEANU, C. M., CAFFO, B. S. and PUNJABI, N. M. (2009). Multilevel functional principal component analysis. *Ann. Appl. Stat.* **3** 458–488. [MR2668715](#) <https://doi.org/10.1214/08-AOAS206>
- FRATI, F., SALVATORI, C., INCORVAIA, C., BELLUCCI, A., DI CARA, G., MARCUCCI, F. and ESPOSITO, S. (2019). The role of the microbiome in asthma: The gut–lung axis. *Int. J. Mol. Sci.* **20** 123.
- GABRY, J., SIMPSON, D., VEHTARI, A., BETANCOURT, M. and GELMAN, A. (2019). Visualization in Bayesian workflow. *J. Roy. Statist. Soc. Ser. A* **182** 389–402. [MR3902665](#) <https://doi.org/10.1111/rssa.12378>
- GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. [MR1422404](#)
- GILL, S. R., POP, M., DEBOY, R. T., ECKBURG, P. B., TURNBAUGH, P. J., SAMUEL, B. S., GORDON, J. I., RELMAN, D. A., FRASER-LIGGETT, C. M. et al. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* **312** 1355–1359.
- HALL, P. and HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 109–126. [MR2212577](#) <https://doi.org/10.1111/j.1467-9868.2005.00535.x>
- HOLLERAN, G., SCALDAFERRI, F., IANIRO, G., LOPETUSO, L., MC, D. N., MELE, M., GASBARRINI, A. and CAMMAROTA, G. (2018). Fecal microbiota transplantation for the treatment of patients with ulcerative colitis and other gastrointestinal conditions beyond *Clostridium difficile* infection: An update. *Drugs Today* **54** 123–136.
- JAMES, G. M., HASTIE, T. J. and SUGAR, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87** 587–602. [MR1789811](#) <https://doi.org/10.1093/biomet/87.3.587>
- JIANG, D., ARMOUR, C. R., HU, C., MEI, M., TIAN, C., SHARPTON, T. J. and JIANG, Y. (2019). Microbiome multiomics network analysis: Statistical considerations, limitations, and opportunities. *Front. Genet.* **10** 995.
- JIANG, L., ZHONG, Y., ELROD, C., NATARAJAN, L., KNIGHT, R. and THOMPSON, W. K. (2020). BayesTime: Bayesian functional principal components for sparse longitudinal data. Preprint. Available at [arXiv:2012.00579](#).
- JIANG, L., ELROD, C., KIM, J. J., SWAFFORD, A. D., KNIGHT, R. and THOMPSON, W. K. (2022). Supplement to “Bayesian multivariate sparse functional principal components analysis with application to longitudinal microbiome multiomics data.” <https://doi.org/10.1214/21-AOAS1587SUPPA>, <https://doi.org/10.1214/21-AOAS1587SUPPB>
- JOE, H. (1989). Relative entropy measures of multivariate dependence. *J. Amer. Statist. Assoc.* **84** 157–164. [MR0999674](#)
- KIDZIŃSKI, Ł. and HASTIE, T. (2018). Longitudinal data analysis using matrix completion. Preprint. Available at [arXiv:1809.08771](#).
- KOSTIC, A. D., GEVERS, D., SILJANDER, H., VATANEN, T., HYÖTYLÄINEN, T., HÄMÄLÄINEN, A.-M., PEET, A., TILLMANN, V., PÖHÖ, P. et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17** 260–273.
- KUCZYŃSKI, J., COSTELLO, E. K., NEMERGUT, D. R., ZANEVELD, J., LAUBER, C. L., KNIGHTS, D., KOREN, O., FIERER, N., KELLEY, S. T. et al. (2010). Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol.* **11** 210.
- LI, Y. and HSING, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.* **38** 3321–3351. [MR2766854](#) <https://doi.org/10.1214/10-AOS813>
- LIECHTY, J. C., LIECHTY, M. W. and MÜLLER, P. (2004). Bayesian correlation estimation. *Biometrika* **91** 1–14. [MR2050456](#) <https://doi.org/10.1093/biomet/91.1.1>
- LLOYD-PRICE, J., ARZE, C., ANANTHAKRISHNAN, A. N., SCHIRMER, M., AVILA-PACHECO, J., POON, T. W., ANDREWS, E., AJAMI, N. J., BONHAM, K. S. et al. (2019). Multiomics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569** 655–662.

- MORTON, J. T., AKSENOV, A. A., NOTHIAS, L. F., FOULDS, J. R., QUINN, R. A., BADRI, M. H., SWENSON, T. L., VAN GOETHEM, M. W., NORTHEN, T. R. et al. (2019). Learning representations of microbe-metabolite interactions. *Nat. Methods* **16** 1306–1314.
- PENG, J. and PAUL, D. (2009). A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *J. Comput. Graph. Statist.* **18** 995–1015. [MR2598035 https://doi.org/10.1198/jcgs.2009.08011](https://doi.org/10.1198/jcgs.2009.08011)
- RAMSAY, J. and SILVERMAN, B. W. (1997). *Functional Data Analysis, Springer Series in Statistics*. Springer, New York. [MR2168993](https://doi.org/10.1198/jcgs.2009.08011)
- RAMSAY, J. O. and SILVERMAN, B. W. (2007). *Applied Functional Data Analysis: Methods and Case Studies*. Springer, New York.
- RANJAN, R., RANI, A., METWALLY, A., MCGEE, H. S. and PERKINS, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* **469** 967–977.
- RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53** 233–243. [MR1094283](https://doi.org/10.1198/jcgs.2009.08011)
- ROSE, S. M. S.-F., CONTREPOIS, K., MONEGHETTI, K. J., ZHOU, W., MISHRA, T., MATARASO, S., DAGAN-ROSENFELD, O., GANZ, A. B., DUNN, J. et al. (2019). A longitudinal big data approach for precision health. *Nat. Med.* **25** 792–804.
- SBERRO, H., FREMIN, B. J., ZLITNI, S., EDFORS, F., GREENFIELD, N., SNYDER, M. P., PAVLOPOULOS, G. A., KYRPIDES, N. C. and BHATT, A. S. (2019). Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell* **178** 1245–1259.
- SHARON, G., CRUZ, N. J., KANG, D.-W., GANDAL, M. J., WANG, B., KIM, Y.-M., ZINK, E. M., CASEY, C. P., TAYLOR, B. C. et al. (2019). Human gut microbiota from autism spectrum disorder promote behavioral symptoms in mice. *Cell* **177** 1600–1618.
- STEWART, C. J., AJAMI, N. J., O'BRIEN, J. L., HUTCHINSON, D. S., SMITH, D. P., WONG, M. C., ROSS, M. C., LLOYD, R. E., DODDAPANENI, H. et al. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562** 583–588.
- VATANEN, T., FRANZOSA, E. A., SCHWAGER, R., TRIPATHI, S., ARTHUR, T. D., VEHIK, K., LERNMARK, Å., HAGOPIAN, W. A., REWERS, M. J. et al. (2018). The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562** 589–594.
- VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27** 1413–1432. [MR3647105 https://doi.org/10.1007/s11222-016-9696-4](https://doi.org/10.1007/s11222-016-9696-4)
- YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. [MR2160561 https://doi.org/10.1198/016214504000001745](https://doi.org/10.1198/016214504000001745)

DATA-ADAPTIVE EFFICIENT ESTIMATION STRATEGIES FOR BIOMARKER STUDIES EMBEDDED IN RANDOMIZED TRIALS

BY WEI ZHANG^{1,a}, ZHIWEI ZHANG^{2,b}, JAMES F. TROENDLE^{3,c} AND AIYI LIU^{4,d}

¹Academy of Mathematics and Systems Science, Chinese Academy of Sciences, ^azhangwei@amss.ac.cn

²National Cancer Institute, National Institutes of Health, ^bzhiwei.zhang@nih.gov

³National Heart, Lung and Blood Institute, National Institutes of Health, ^cjames.troendle@nih.gov

⁴Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, ^dliua@mail.nih.gov

Predictive and prognostic biomarkers are increasingly important in clinical research and practice. Biomarker studies are frequently embedded in randomized clinical trials with biospecimens collected at baseline and assayed for biomarkers, either in real time or retrospectively. This article proposes efficient estimation strategies for two study settings in terms of biomarker ascertainment: a complete-data setting in which the biomarker is measured for all subjects in the trial, and a two-phase sampling design in which the biomarker is measured retrospectively for a random subsample of subjects selected in an outcome-dependent fashion. In both settings, efficient estimating functions are characterized using semiparametric theory and approximated using data-adaptive machine learning methods, leading to estimators that are consistent, asymptotically normal and (approximately) efficient under general conditions. The proposed methods are evaluated in simulation studies and applied to real data from two biomarker studies, one in each setting.

REFERENCES

- BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins Univ. Press, Baltimore, MD. [MR1245941](#)
- CARROLL, R. J. and WAND, M. P. (1991). Semiparametric estimation in logistic measurement error models. *J. Roy. Statist. Soc. Ser. B* **53** 573–585. [MR1125715](#)
- CHATTERJEE, N., CHEN, Y.-H. and BRESLOW, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *J. Amer. Statist. Assoc.* **98** 158–168. [MR1965682](#) <https://doi.org/10.1198/016214503388619184>
- CHEN, H. Y. (2004). Nonparametric and semiparametric models for missing covariates in parametric regression. *J. Amer. Statist. Assoc.* **99** 1176–1189. [MR2109505](#) <https://doi.org/10.1198/016214504000001727>
- CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. [MR3769544](#) <https://doi.org/10.1111/ectj.12097>
- DURIE, B. G. M., HOERING, A., ABIDI, M. H., RAJKUMAR, S. V., EPSTEIN, J., KAHANIC, S. P., THAKURI, M., REU, F., REYNOLDS, C. M. et al. (2017). Bortezomib with lenalidomide and dexamethasone versus lenalidomide and dexamethasone alone in patients with newly diagnosed myeloma without intent for immediate autologous stem-cell transplant (SWOG S0777): A randomised, open-label, phase 3 trial. *Lancet* **389** 519–527. [https://doi.org/10.1016/S0140-6736\(16\)31594-X](https://doi.org/10.1016/S0140-6736(16)31594-X)
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer Series in Statistics. Springer, New York. [MR2722294](#) <https://doi.org/10.1007/978-0-387-84858-7>
- IBRAHIM, J. G., CHEN, M. and LIPSITZ, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics* **55** 591–596.
- KALFF, A. and SPENCER, A. (2012). The $t(4; 14)$ translocation and FGFR3 overexpression in multiple myeloma: Prognostic implications and current clinical strategies. *Blood Cancer J.* **2**.

Key words and phrases. Augmentation, precision medicine, semiparametric theory, super learner, two-phase sampling.

- KARAPETIS, C. S., KHAMBATA-FORD, S., JONKER, D. J., O'CALLAGHAN, C. J., TU, D., TEBBUTT, N. C., SIMES, R. J., CHALCHAL, H., SHAPIRO, J. D. et al. (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N. Engl. J. Med.* **359** 1757–1765.
- KENNEDY, E. H., BALAKRISHNAN, S. and G'SELL, M. (2020). Sharp instruments for classifying compliers and generalizing causal effects. *Ann. Statist.* **48** 2008–2030. [MR4134784 https://doi.org/10.1214/19-AOS1874](https://doi.org/10.1214/19-AOS1874)
- KOOPERBERG, C., CUSHMAN, M., HSIA, J., ROBINSON, J. G., ARAFAKI, A. K., LYNCH, J. K., BAIRD, A. E., JOHNSON, K. C., KULLER, L. H. et al. (2007). Can biomarkers identify women at increased stroke risk? The women's health initiative hormone trials. *PLOS Clin. Trials* **2**.
- PEPE, M. S. and FLEMING, T. R. (1991). A nonparametric method for dealing with mismeasured covariate data. *J. Amer. Statist. Assoc.* **86** 108–113. [MR1137103](https://doi.org/10.1080/01621459.2014.951443)
- POLLEY, E. C., ROSE, S. and VAN DER LAAN, M. J. (2011). Super learning. In *Targeted Learning*. Springer Ser. Statist. 43–66. Springer, New York. [MR2867115 https://doi.org/10.1007/978-1-4419-9782-1_3](https://doi.org/10.1007/978-1-4419-9782-1_3)
- POLLEY, M. Y. C., FREIDLIN, B., KORN, E. L., CONLEY, B. A., ABRAMS, J. S. and MCSHANE, L. M. (2013). Statistical and practical considerations for clinical evaluation of predictive biomarkers. *J. Natl. Cancer Inst.* **105** 1677–1683.
- ROSSOUW, J. E., ANDERSON, G. L., PRENTICE, R. L., LACROIX, A. Z., KOOPERBERG, C., STEFANICK, M. L., JACKSON, R. D., BERESFORD, S. A. A., HOWARD, B. V. et al. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the women's health initiative randomized controlled trial. *J. Am. Med. Assoc.* **288** 321–333.
- RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. [MR0455196 https://doi.org/10.1093/biomet/63.3.581](https://doi.org/10.1093/biomet/63.3.581)
- SIMES, J., ROBLEDO, K. P., WHITE, H. D., ESPINOZA, D., STEWART, R. A., SULLIVAN, D. R., ZELLER, T., HAGUE, W., NESTEL, P. J. et al. (2018). D-dimer predicts long-term cause-specific mortality, cardiovascular events, and cancer in patients with stable coronary heart disease. *Circulation* **138** 721–723.
- SIMON, R., PAIK, S. and HAYES, D. F. (2009). Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J. Natl. Cancer Inst.* **101** 1446–1452.
- TIAN, L., ALIZADEH, A. A., GENTLES, A. J. and TIBSHIRANI, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *J. Amer. Statist. Assoc.* **109** 1517–1532. [MR3293607 https://doi.org/10.1080/01621459.2014.951443](https://doi.org/10.1080/01621459.2014.951443)
- TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer, New York. [MR2233926](https://doi.org/10.1002/sim.3113)
- TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LU, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Stat. Med.* **27** 4658–4677. [MR2528575 https://doi.org/10.1002/sim.3113](https://doi.org/10.1002/sim.3113)
- VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer, New York. [MR1958123 https://doi.org/10.1007/978-0-387-21700-0](https://doi.org/10.1007/978-0-387-21700-0)
- VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics **3**. Cambridge Univ. Press, Cambridge. [MR1652247 https://doi.org/10.1017/CBO9780511802256](https://doi.org/10.1017/CBO9780511802256)
- VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer Series in Statistics. Springer, New York. [MR1385671 https://doi.org/10.1007/978-1-4757-2545-2](https://doi.org/10.1007/978-1-4757-2545-2)
- ZHANG, Z. and MA, S. (2019). Machine learning methods for leveraging baseline covariate information to improve the efficiency of clinical trials. *Stat. Med.* **38** 1703–1714. [MR3934815 https://doi.org/10.1002/sim.8054](https://doi.org/10.1002/sim.8054)
- ZHANG, M., TSIATIS, A. A. and DAVIDIAN, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* **64** 707–715. [MR2526620 https://doi.org/10.1111/j.1541-0420.2007.00976.x](https://doi.org/10.1111/j.1541-0420.2007.00976.x)
- ZHANG, W., ZHANG, Z. W., TROENDLE, J. F. and LIU, A. (2022). Supplement to “Data-adaptive efficient estimation strategies for biomarker studies embedded in randomized trials.” <https://doi.org/10.1214/21-AOAS1588SUPP>
- ZHANG, Z., QU, Y., ZHANG, B., NIE, L. and SOON, G. (2016). Use of auxiliary covariates in estimating a biomarker-adjusted treatment effect model with clinical trial data. *Stat. Methods Med. Res.* **25** 2103–2119. [MR3553329 https://doi.org/10.1177/0962280213515572](https://doi.org/10.1177/0962280213515572)
- ZHENG, W. and VAN DER LAAN, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*. Springer Ser. Statist. 459–474. Springer, New York. [MR2867139 https://doi.org/10.1007/978-1-4419-9782-1_27](https://doi.org/10.1007/978-1-4419-9782-1_27)

AN OMNIBUS TEST FOR DETECTION OF SUBGROUP TREATMENT EFFECTS VIA DATA PARTITIONING

BY YIFEI SUN^{1,a}, XUMING HE^{2,c} AND JIANHUA HU^{1,b}

¹Department of Biostatistics, Columbia University, ^ays3072@cumc.columbia.edu, ^bjh3992@cumc.columbia.edu

²Department of Statistics, University of Michigan, ^cxmhe@umich.edu

Late-stage clinical trials have been conducted primarily to establish the efficacy of a new treatment in an intended population. A corollary of population heterogeneity in clinical trials is that a treatment might be effective for one or more subgroups, rather than for the whole population of interest. As an example, the phase III clinical trial of panitumumab in metastatic colorectal cancer patients failed to demonstrate its efficacy in the overall population, but a subgroup associated with tumor KRAS status was found to be promising (Peeters et al. (*Am. J. Clin. Oncol.* **28** (2010) 4706–4713)). As we search for such subgroups via data partitioning based on a large number of biomarkers, we need to guard against inflated type I error rates due to multiple testing. Commonly-used multiplicity adjustments tend to lose power for the detection of subgroup treatment effects. We develop an effective omnibus test to detect the existence of, at least, one subgroup treatment effect, allowing a large number of possible subgroups to be considered and possibly censored outcomes. Applied to the panitumumab trial data, the proposed test would confirm a significant subgroup treatment effect. Empirical studies also show that the proposed test is applicable to a variety of outcome variables and maintains robust statistical power.

REFERENCES

- BEHR, M., ANSARI, M. A., MUNK, A. and HOLMES, C. (2020). Testing for dependence on tree structures. *Proc. Natl. Acad. Sci. USA* **117** 9787–9792. [MR4236178](#) <https://doi.org/10.1073/pnas.1912957117>
- CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. [MR3161448](#) <https://doi.org/10.1214/13-AOS1161>
- CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* **45** 2309–2352. [MR3693963](#) <https://doi.org/10.1214/16-AOP1113>
- CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2019). Inference on causal and structural parameters using many moment inequalities. *Rev. Econ. Stud.* **86** 1867–1900. [MR4009488](#) <https://doi.org/10.1093/restud/rdy065>
- DMITRIENKO, A., MILLEN, B. and LIPKOVICH, I. (2017). Multiplicity considerations in subgroup analysis. *Stat. Med.* **36** 4446–4454. [MR3731226](#) <https://doi.org/10.1002/sim.7416>
- DMITRIENKO, A., LIPKOVICH, I., DANE, A. and MUYSERS, C. (2020). Data-driven and confirmatory subgroup analysis in clinical trials. In *Design and Analysis of Subgroups with Biopharmaceutical Applications* (N. Ting, J. C. Cappelleri, S. Ho and D. Chen, eds.) 33–91. Springer, Berlin.
- GEHAN, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52** 203–223. [MR0207130](#) <https://doi.org/10.1093/biomet/52.1-2.203>
- GUO, X. and HE, X. (2021). Inference on selected subgroups in clinical trials. *J. Amer. Statist. Assoc.* **116** 1498–1506. [MR4309288](#) <https://doi.org/10.1080/01621459.2020.1740096>
- HUANG, Y. (2013). Fast censored linear regression. *Scand. J. Stat.* **40** 789–806. [MR3145118](#) <https://doi.org/10.1111/sjso.12031>
- JOSHI, N., FINE, J., CHU, R. and IVANOVA, A. (2019). Estimating the subgroup and testing for treatment effect in a post-hoc analysis of a clinical trial with a biomarker. *J. Biopharm. Statist.* **29** 685–695.
- LEE, J. J. and RUBIN, D. B. (2016). Evaluating the validity of post-hoc subgroup inferences: A case study. *Amer. Statist.* **70** 39–46. [MR3480669](#) <https://doi.org/10.1080/00031305.2015.1093961>

Key words and phrases. Bootstrap, data partitioning, clinical trials, high-dimensional covariates, subgroup treatment effect.

- LOH, W.-Y. (2009). Improving the precision of classification trees. *Ann. Appl. Stat.* **3** 1710–1737. [MR2752155](https://doi.org/10.1214/09-AOAS260) <https://doi.org/10.1214/09-AOAS260>
- LOH, W.-Y., HE, X. and MAN, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Stat. Med.* **34** 1818–1833. [MR3334694](https://doi.org/10.1002/sim.6454) <https://doi.org/10.1002/sim.6454>
- LU, W., GOLDBERG, Y. and FINE, J. P. (2012). On the robustness of the adaptive lasso to model misspecification. *Biometrika* **99** 717–731. [MR2966780](https://doi.org/10.1093/biomet/ass027) <https://doi.org/10.1093/biomet/ass027>
- NAGGARA, O., RAYMOND, J., GUILBERT, F. and ALTMAN, D. G. (2011). The problem of subgroup analyses: An example from a trial on ruptured intracranial aneurysms. *Am. J. Neuroradiol.* **32** 633–636.
- O'BRIEN, T. R., KUHS, K. A. and PFEIFFER, R. M. (2014). Subgroup differences in response to 8 weeks of ledipasvir/sofosbuvir for chronic hepatitis C. *Open Forum Infect. Dis.* **1** ofu110.
- ONDRA, T., DMITRIENKO, A., FRIEDE, T., GRAF, A., MILLER, F., STALLARD, N. and POSCH, M. (2016). Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *J. Biopharm. Statist.* **26** 99–119. <https://doi.org/10.1080/10543406.2015.1092034>
- PEETERS, M., PRICE, T., CERVANTES, A., SOBRERO, A., DUCREUX, M., HOTKO, Y., ANDRÉ, T., CHAN, E., LORDICK, F. et al. (2010). Randomized phase III study of panitumumab with fluorouracil, leucovorin, and irinotecan (FOLFIRI) compared with FOLFIRI alone as second-line treatment in patients with metastatic colorectal cancer. *Am. J. Clin. Oncol.* **28** 4706–4713.
- PEETERS, M., OLINER, K. S., PRICE, T. J., CERVANTES, A., SOBRERO, A. F., DUCREUX, M., HOTKO, Y., ANDRÉ, T., CHAN, E. et al. (2015). Analysis of KRAS/NRAS mutations in a phase III study of panitumumab with FOLFIRI compared with FOLFIRI alone as second-line treatment for metastatic colorectal cancer. *Clin. Cancer Res.* **21** 5469–5479.
- PEPE, M. S. and FLEMING, T. R. (1991). Weighted Kaplan-Meier statistics: Large sample and optimality considerations. *J. Roy. Statist. Soc. Ser. B* **53** 341–352. [MR1108331](https://doi.org/10.1080/01621459.2014.894763)
- SHEN, J. and HE, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *J. Amer. Statist. Assoc.* **110** 303–312. [MR3338504](https://doi.org/10.1080/01621459.2014.894763) <https://doi.org/10.1080/01621459.2014.894763>
- SHEN, J. and QU, A. (2020). Subgroup analysis based on structured mixed-effects models for longitudinal data. *J. Biopharm. Statist.* **30** 607–622.
- SIES, A., DEMYTTENAERE, K. and MECHELEN, I. V. (2019). Studying treatment-effect heterogeneity in precision medicine through induced subgroups. *J. Biopharm. Statist.* **29** 491–507.
- SONG, R., LU, W., MA, S. and JENG, X. J. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika* **101** 799–814. [MR3286918](https://doi.org/10.1093/biomet/asu047) <https://doi.org/10.1093/biomet/asu047>
- SUN, Y., HE, X. and HU, J. (2022). Supplement to “An omnibus test for detection of subgroup treatment effects via data partitioning.” <https://doi.org/10.1214/21-AOAS1589SUPP>
- TSIATIS, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18** 354–372. [MR1041397](https://doi.org/10.1214/aos/1176347504) <https://doi.org/10.1214/aos/1176347504>
- 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) <https://doi.org/10.1080/01621459.2017.1319839>
- ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. [MR2279469](https://doi.org/10.1198/016214506000000735) <https://doi.org/10.1198/016214506000000735>

ASSESSING TREATMENT EFFECT THROUGH COMPLIANCE SCORE IN RANDOMIZED TRIALS WITH NONCOMPLIANCE

BY ZONGHUI HU^{1,a}, ZHIWEI ZHANG^{2,c} AND DEAN FOLLMANN^{1,b}

¹National Institute of Allergy and Infectious Diseases, ^ahuzo@niaid.nih.gov, ^bdfollmann@niaid.nih.gov

²National Cancer Institute, ^czhiwei.zhang@nih.gov

A randomized trial is the gold standard for assessing the benefit of a treatment versus a control. When noncompliance is present, treatment effect depends on the tendency to comply—an attribute that is not directly measurable. Though the principal causal effect has been the most important for handling noncompliance, it is not immediately applicable to clinical decision-making as it targets the average effect in the latent strata of potential compliance. In this work, we propose the concept of compliance score, a linear combination of baseline characteristics, that uncovers the inherent attribute of compliance. We then assess the heterogeneous causal effect, namely, the causal effect of treatment as a function of baseline characteristics through the compliance score. A pseudo-response, along with a nonparametric estimation procedure, is proposed to ensure consistent and optimally efficient estimation. Compare to principal causal effect, the proposed effect is actionable and allows prediction of treatment effect at individual level. This work is motivated by and applied to a clinical trial to evaluate the benefit of antiretroviral regimens in HIV-infected patients.

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](https://doi.org/10.1016/S0304-4076(02)00201-4)
- BARNARD, J., FRANGAKIS, C. E., HILL, J. L. and RUBIN, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *J. Amer. Statist. Assoc.* **98** 299–323. [MR1995712 https://doi.org/10.1198/0162145030000071](https://doi.org/10.1198/0162145030000071)
- CHIAROMONTE, F., COOK, R. D. and LI, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *Ann. Statist.* **30** 475–497. [MR1902896 https://doi.org/10.1214/aos/1021379862](https://doi.org/10.1214/aos/1021379862)
- COOK, R. D. and LEE, H. (1999). Dimension reduction in binary response regression. *J. Amer. Statist. Assoc.* **94** 1187–1200. [MR1731482 https://doi.org/10.2307/2669934](https://doi.org/10.2307/2669934)
- COOK, R. D. and LI, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30** 455–474. [MR1902895 https://doi.org/10.1214/aos/1021379861](https://doi.org/10.1214/aos/1021379861)
- CUI, X., HÄRDLE, W. K. and ZHU, L. (2011). The EFM approach for single-index models. *Ann. Statist.* **39** 1658–1688. [MR2850216 https://doi.org/10.1214/10-AOS871](https://doi.org/10.1214/10-AOS871)
- DING, P. and LU, J. (2017). Principal stratification analysis using principal scores. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 757–777. [MR3641406 https://doi.org/10.1111/rssb.12191](https://doi.org/10.1111/rssb.12191)
- DUNN, G., MARACY, M. and TOMENSON, B. (2005). Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: The role of instrumental variable methods. *Stat. Methods Med. Res.* **14** 369–395. [MR2201833 https://doi.org/10.1191/0962280205sm403oa](https://doi.org/10.1191/0962280205sm403oa)
- EFRON, B. and FELDMAN, D. (1991). Compliance as an explanatory variable in clinical trials. *J. Amer. Statist. Assoc.* **86** 9–17.
- FAN, J. (2018). *Local Polynomial Modelling and Its Applications. Monographs on Statistics and Applied Probability* **66**. Routledge, London.
- FELLER, A., MEALLI, F. and MIRATRIX, L. (2017). Principal score methods: Assumptions, extensions, and practical considerations. *J. Educ. Behav. Stat.* **42** 726–758.
- FOLLMANN, D. A. (2000). On the effect of treatment among would-be treatment compliers: An analysis of the multiple risk factor intervention trial. *J. Amer. Statist. Assoc.* **95** 1101–1109. [MR1821718 https://doi.org/10.2307/2669746](https://doi.org/10.2307/2669746)

Key words and phrases. Causal inference, noncompliance, nonparametric regression, principal causal effect, principal stratification causal effect, randomized trial.

- FRANGAKIS, C. E. and RUBIN, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86** 365–379. [MR1705410 https://doi.org/10.1093/biomet/86.2.365](https://doi.org/10.1093/biomet/86.2.365)
- FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. [MR1891039 https://doi.org/10.1111/j.0006-341X.2002.00021.x](https://doi.org/10.1111/j.0006-341X.2002.00021.x)
- HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. [MR1212171 https://doi.org/10.1214/aos/1176349020](https://doi.org/10.1214/aos/1176349020)
- HASTIE, T. J. (2017). *Generalized Additive Models*. Routledge, London.
- HERNÁN, M. A. and HERNÁNDEZ-DÍAZ, S. (2012). Beyond the intention-to-treat in comparative effectiveness research. *Clin. Trials* **9** 48–55.
- HU, Z., ZHANG, Z. and FOLLMANN, D. (2022). Supplement to “Assessing treatment effect through compliance score in randomized trials with noncompliance.” <https://doi.org/10.1214/21-AOAS1590SUPP>
- ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58** 71–120. [MR1230981 https://doi.org/10.1016/0304-4076\(93\)90114-K](https://doi.org/10.1016/0304-4076(93)90114-K)
- JIN, H. and RUBIN, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *J. Amer. Statist. Assoc.* **103** 101–111. [MR2463484 https://doi.org/10.1198/016214507000000347](https://doi.org/10.1198/016214507000000347)
- JOFFE, M. M., HAVE, T. R. T. and BRENSINGER, C. (2003). The compliance score as a regressor in randomized trials. *Biostatistics* **4** 327–340.
- LI, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R. Monographs on Statistics and Applied Probability* **161**. CRC Press, Boca Raton, FL. [MR3838449 https://doi.org/10.1201/9781315119427](https://doi.org/10.1201/9781315119427)
- PEARL, J. (2011). Principal stratification—a goal or a tool? *Int. J. Biostat.* **7** Art. 20, 15. [MR2787410 https://doi.org/10.2202/1557-4679.1322](https://doi.org/10.2202/1557-4679.1322)
- PHIDISA II WRITING TEAM FOR PROJECT PHIDISA, RATSELA, A., POLIS, M., DHLOMO, S., EMERY, S., GRANDITS, G., KHABO, P., KHANYILE, T., KOMATI, S. et al. (2010). A randomized factorial trial comparing 4 treatment regimens in treatment-naïve HIV-infected persons with AIDS and/or a CD4 cell count <200 cells/ μ l in South Africa. *J. Infect. Dis.* **202** 1529–1537. <https://doi.org/10.1086/656718>
- ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 143–155.
- ROSENBAUM, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Stat. Soc., A* **147** 656–666.
- TAN, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *J. Amer. Statist. Assoc.* **101** 1607–1618. [MR2279483 https://doi.org/10.1198/016214505000001366](https://doi.org/10.1198/016214505000001366)
- TAN, Z. (2010). Marginal and nested structural models using instrumental variables. *J. Amer. Statist. Assoc.* **105** 157–169. [MR2757199 https://doi.org/10.1198/jasa.2009.tm08299](https://doi.org/10.1198/jasa.2009.tm08299)
- VANDERWEELE, T. J. (2011). Principal stratification—uses and limitations. *Int. J. Biostat.* **7** Art. 28, 14. [MR2843532 https://doi.org/10.2202/1557-4679.1329](https://doi.org/10.2202/1557-4679.1329)
- XIA, Y., TONG, H., LI, W. K. and ZHU, L.-X. (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 363–410. [MR1924297 https://doi.org/10.1111/1467-9868.03411](https://doi.org/10.1111/1467-9868.03411)

SPATIOTEMPORAL SATELLITE DATA IMPUTATION USING SPARSE FUNCTIONAL DATA ANALYSIS

BY WEICHENG ZHU^a, ZHENGYUAN ZHU^b AND XIONGTAO DAI^c

Department of Statistics, Iowa State University, ^aweichzhu@amazon.com, ^bzhuz@iastate.edu, ^cxdai@iastate.edu

Many scientific applications and signal processing algorithms require complete satellite images. However, missing data in satellite images is very common due to various reasons such as cloud cover and sensor-specific problems. This paper introduces a general spatiotemporal satellite image imputation method based on sparse functional data analytic techniques. To handle observations consisting of a few longitudinally repeated satellite images that are themselves partially observed and noise-contaminated, we propose a multistep imputation method by following the best linear unbiased prediction principle and pooling information across all available locations and time points. Theoretical properties are established for the proposed approach under a new observation model for functional data that covers the dataset in question as a special case. Practical analysis on the Landsat data are conducted to illustrate and validate our algorithm which also shows that the proposed method considerably outperforms existing algorithms in terms of prediction accuracy. An efficient implementation using R and Rcpp is made available in the R package `stfit`.

REFERENCES

- CHEN, L.-H. and JIANG, C.-R. (2017). Multi-dimensional functional principal component analysis. *Stat. Comput.* **27** 1181–1192. [MR3647091 https://doi.org/10.1007/s11222-016-9679-5](https://doi.org/10.1007/s11222-016-9679-5)
- CHEN, J., ZHU, X., VOGELMANN, J. E., GAO, F. and JIN, S. (2011). A simple and effective method for filling gaps in landsat etm+ slc-off images. *Remote Sens. Environ.* **115** 1053–1064.
- CRESSIE, N. and WIKLE, C. K. (2015). *Statistics for Spatio-Temporal Data*. Wiley, New York.
- GERBER, F., DE JONG, R., SCHAEPMAN, M. E., SCHAEPMAN-STRUB, G. and FURRER, R. (2018). Predicting missing values in spatio-temporal remote sensing data. *IEEE Trans. Geosci. Remote Sens.* **56** 2841–2853.
- GNEITING, T., GENTON, M. G. and GUTTORP, P. (2006). Geostatistical space-time models, stationarity, separability, and full symmetry. *Monogr. Statist. Appl. Probab.* **107** 151.
- JAMES, G. M., HASTIE, T. J. and SUGAR, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87** 587–602. [MR1789811 https://doi.org/10.1093/biomet/87.3.587](https://doi.org/10.1093/biomet/87.3.587)
- JÖNSSON, P. and EKLUNDH, L. (2004). Timesat—a program for analyzing time-series of satellite sensor data. *Comput. Geosci.* **30** 833–845.
- KATZFUSS, M. and CRESSIE, N. (2011). Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets. *J. Time Series Anal.* **32** 430–446. [MR2841794 https://doi.org/10.1111/j.1467-9892.2011.00732.x](https://doi.org/10.1111/j.1467-9892.2011.00732.x)
- KING, M. D., PLATNICK, S., MENZEL, W. P., ACKERMAN, S. A. and HUBANKS, P. A. (2013). Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites. *IEEE Trans. Geosci. Remote Sens.* **51** 3826–3852.
- KRAUS, D. (2015). Components and completion of partially observed functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 777–801. [MR3382597 https://doi.org/10.1111/rssb.12087](https://doi.org/10.1111/rssb.12087)
- LI, Y. and HSING, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.* **38** 3321–3351. [MR2766854 https://doi.org/10.1214/10-AOS813](https://doi.org/10.1214/10-AOS813)
- MA, P. and KANG, E. L. (2020). A fused Gaussian process model for very large spatial data. *J. Comput. Graph. Statist.* **29** 479–489. [MR4153175 https://doi.org/10.1080/10618600.2019.1704293](https://doi.org/10.1080/10618600.2019.1704293)
- MERSMANN, O. (2019). `microbenchmark`: Accurate Timing Functions. R package version 1.4-7.
- MICROSOFT and WESTON, S. (2020). `foreach`: Provides Foreach Looping Construct. R package version 1.5.0.

- MONTERO, J.-M., FERNÁNDEZ-AVILÉS, G. and MATEU, J. (2015). *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging* **998**. Wiley, New York.
- MORENO, Á., GARCÍA-HARO, F. J., MARTÍNEZ, B. and GILABERT, M. A. (2014). Noise reduction and gap filling of fapar time series using an adapted local regression filter. *Remote Sens.* **6** 8238–8260.
- MÜLLER, H.-G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *Ann. Statist.* **12** 766–774. [MR0740929 https://doi.org/10.1214/aos/1176346523](https://doi.org/10.1214/aos/1176346523)
- NASA (2016). Landsat 7 science data users handbook. Available at https://landsat.gsfc.nasa.gov/wp-content/uploads/2016/08/Landsat7_Handbook.pdf. [Online; accessed 13-May-2019].
- NASA (2019). Landsat applications. Available at <https://landsat.gsfc.nasa.gov/categories/applications>. [Online; accessed 12-Sep-2021].
- NUSSER, S., BREIDT, F. and FULLER, W. (1998). Design and estimation for investigating the dynamics of natural resources. *Ecol. Appl.* **8** 234–245.
- NUSSER, S. M. and GOEBEL, J. J. (1997). The national resources inventory: A long-term multi-resource monitoring programme. *Environ. Ecol. Stat.* **4** 181–204.
- POGGIO, L., GIMONA, A. and BROWN, I. (2012). Spatio-temporal modis evi gap filling under cloud cover: An example in Scotland. *ISPRS J. Photogramm. Remote Sens.* **72** 56–72.
- PRINGLE, M., SCHMIDT, M. and MUIR, J. (2009). Geostatistical interpolation of slc-off landsat etm+ images. *ISPRS J. Photogramm. Remote Sens.* **64** 654–664.
- RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR2168993](https://doi.org/10.1007/978-1-4939-9738-9)
- RIBEIRO, P. J. JR and DIGGLE, P. J. (2016). *GeoR: Analysis of Geostatistical Data*. R package version 1.7-5.2.
- ROERINK, G., MENENTI, M. and VERHOEF, W. (2000). Reconstructing cloudfree ndvi composites using Fourier analysis of time series. *Int. J. Remote Sens.* **21** 1911–1917.
- STEIN, M. L. (2005). Space-time covariance functions. *J. Amer. Statist. Assoc.* **100** 310–321. [MR2156840 https://doi.org/10.1198/016214504000000854](https://doi.org/10.1198/016214504000000854)
- USDA (2015). *Summary Report: 2012 National Resources Inventory*. Natural Resources Conservation Service, Washington, DC, and Center for Survey Statistics and Methodology. Iowa State Univ., Ames, IA. Available at https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcseprd396218.pdf. [Online; accessed 05-July-2018].
- USGS (2004). Phase 2 gap-fill algorithm: Slc-off gap-filled products gap-fill algorithm methodology. Accessed: 2017-11-26.
- VERGER, A., BARET, F., WEISS, M., KANDASAMY, S. and VERMOTE, E. (2013). The cacao method for smoothing, gap filling, and characterizing seasonal anomalies in satellite time series. *IEEE Trans. Geosci. Remote Sens.* **51** 1963–1972.
- WEISS, D. J., ATKINSON, P. M., BHATT, S., MAPPIN, B., HAY, S. I. and GETHING, P. W. (2014). An effective approach for gap-filling continental scale remotely sensed time-series. *ISPRS J. Photogramm. Remote Sens.* **98** 106–118. <https://doi.org/10.1016/j.isprsjprs.2014.10.001>
- WIKIPEDIA (2018). Landsat 7. Available at https://en.wikipedia.org/wiki/Landsat_7. [Online; accessed 25-June-2018].
- YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. [MR2160561 https://doi.org/10.1198/016214504000001745](https://doi.org/10.1198/016214504000001745)
- ZENG, Z., LEI, L., HOU, S., RU, F., GUAN, X. and ZHANG, B. (2014). A regional gap-filling method based on spatiotemporal variogram model of CO₂ columns. *IEEE Trans. Geosci. Remote Sens.* **52** 3594–3603.
- ZHANG, C., LI, W. and TRAVIS, D. (2007). Gaps-fill of slc-off landsat etm+ satellite image using a geostatistical approach. *Int. J. Remote Sens.* **28** 5103–5122.
- ZHANG, B., SANG, H. and HUANG, J. Z. (2015). Full-scale approximations of spatio-temporal covariance models for large datasets. *Statist. Sinica* **25** 99–114. [MR3328805](https://doi.org/10.1007/s11464-015-0446-4)
- ZHANG, X. and WANG, J.-L. (2016). From sparse to dense functional data and beyond. *Ann. Statist.* **44** 2281–2321. [MR3546451 https://doi.org/10.1214/16-AOS1446](https://doi.org/10.1214/16-AOS1446)
- ZHU, W. (2020). stfit: Spatio-Temporal Functional Imputation Tool. R package version 0.99.8.
- ZHU, W., ZHU, Z. and DAI, X. (2022a). Supplementary material A for “Spatiotemporal satellite data imputation based on sparse functional data analysis.” <https://doi.org/10.1214/21-AOAS1591SUPPA>
- ZHU, W., ZHU, Z. and DAI, X. (2022b). Supplementary material B for “Spatiotemporal satellite data imputation based on sparse functional data analysis.” <https://doi.org/10.1214/21-AOAS1591SUPPB>
- ZHU, W., ZHU, Z. and DAI, X. (2022c). Supplement to “Spatiotemporal satellite data imputation using sparse functional data analysis.” <https://doi.org/10.1214/21-AOAS1591SUPPC>

MULTILEVEL TIME-SERIES MODELS FOR SMALL AREA ESTIMATION AT DIFFERENT FREQUENCIES AND DOMAIN LEVELS

BY HARM JAN BOONSTRA^{1,a} AND JAN VAN DEN BRAKEL^{2,b}

¹*Statistics Netherlands, hhta@cbs.nl*

²*Maastricht University, jbri@cbs.nl*

A small area estimation method is developed for repeatedly conducted multipurpose surveys. A multilevel time-series model is proposed that uses direct estimates for the most detailed domains observed at the highest frequency of the repeated survey. A consistent set of estimates at different aggregation levels is then derived by aggregation of the model-based predictions obtained for the most detailed domains observed at the highest frequency. The model borrows strength over time and space via smooth and local level trends at different aggregation levels. The model also borrows information from auxiliary series available from registers with coefficients that can vary over both domains and time. Regional domain random effects are allowed to vary smoothly over space according to a spatial autoregressive process. To account for the diversity of domains and for more volatile time-dependence, nonnormally distributed random effects and trend innovations are used via so-called global-local shrinkage priors. A Bayesian approach is taken, and the model is estimated by MCMC simulation. The method is illustrated with an application to the Dutch Labour Force Survey to produce monthly provincial and quarterly municipal unemployment figures.

REFERENCES

- ANDREWS, D. F. and MALLOWS, C. L. (1974). Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B* **36** 99–102. [MR0359122](#)
- BATTESE, G., HARTE, R. and FULLER, W. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.* **83** 28–36.
- BELL, W. R., DATTA, G. S. and GHOSH, M. (2013). Benchmarking small area estimators. *Biometrika* **100** 189–202. [MR3034332](#) <https://doi.org/10.1093/biomet/ass063>
- BESAG, J. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746. [MR1380811](#) <https://doi.org/10.1093/biomet/82.4.733>
- BESAG, J., YORK, J. and MOLLIE, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43** 1–20. [MR1105822](#) <https://doi.org/10.1007/BF00116466>
- BOONSTRA, H. J. (2020). *mcmcscsae*: Markov Chain Monte Carlo Small Area Estimation. R package version 0.5.0. <http://CRAN.R-project.org/package=mcmcscsae>.
- BOONSTRA, H. J. and VAN DEN BRAKEL, J. (2019). Estimation of level and change for unemployment using multilevel and structural time-series models. *Surv. Methodol.* **45** 395–425.
- BOONSTRA, H. J. and VAN DEN BRAKEL, J. (2022a). Supplement to “Multilevel time-series models for small area estimation at different frequencies and domain levels.” <https://doi.org/10.1214/21-AOAS1592SUPPA>
- BOONSTRA, H. J. and VAN DEN BRAKEL, J. (2022b). Supplement to “Multilevel time-series models for small area estimation at different frequencies and domain levels.” <https://doi.org/10.1214/21-AOAS1592SUPPB>
- BREZGER, A., FAHRMEIR, L. and HENNERFEIND, A. (2007). Adaptive Gaussian Markov random fields with applications in human brain mapping. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **56** 327–345. [MR2370993](#) <https://doi.org/10.1111/j.1467-9876.2007.00580.x>
- CARTER, C. K. and KOHN, R. (1996). Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika* **83** 589–601. [MR1423876](#) <https://doi.org/10.1093/biomet/83.3.589>
- 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>

- CHUNG, H. C. and DATTA, G. (2020). Bayesian hierarchical spatial models for small area estimation, Technical Report Research report series 2020-07, U.S. Census Bureau, Washington. Available at <https://www.census.gov/library/working-papers/2020/adrm/RRS2020-07.html>.
- DATTA, G. S. and LAHIRI, P. (1995). Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers. *J. Multivariate Anal.* **54** 310–328. [MR1345542](#) <https://doi.org/10.1006/jmva.1995.1059>
- DATTA, G., LAHIRI, P., MAITI, T. and LU, K. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *J. Amer. Statist. Assoc.* **94** 1074–1082.
- FABRIZI, E., FERRANTE, M. R. and TRIVISANO, C. (2018). Bayesian small area estimation for skewed business survey variables. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 861–879. [MR3832254](#) <https://doi.org/10.1111/rssc.12254>
- FABRIZI, E. and TRIVISANO, C. (2010). Robust linear mixed models for small area estimation. *J. Statist. Plann. Inference* **140** 433–443. [MR2558375](#) <https://doi.org/10.1016/j.jspi.2009.07.022>
- FAHRMEIR, L., KNEIB, T. and LANG, S. (2004). Penalized structured additive regression for space–time data: A Bayesian perspective. *Statist. Sinica* **14** 731–761. [MR2087971](#)
- FAY, R. E. and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277. [MR0548019](#)
- GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. [MR1141740](#)
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. [MR2221284](#) <https://doi.org/10.1214/06-BA117A>
- GELMAN, A. and HILL, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge Univ. Press, Cambridge.
- GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.
- HARVEY, A. (2006). Seasonality and unobserved components models: an overview, Technical Report ISSN 1725-4825, Conference on seasonality, seasonal adjustment and their implications for short-term analysis and forecasting, 10-12 May 2006, Luxembourg.
- HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. CRC Press, London. [MR1082147](#)
- LANG, S., FRONK, E.-M. and FAHRMEIR, L. (2002). Function estimation with locally adaptive dynamic models. *Comput. Statist.* **17** 479–499. [MR1952693](#) <https://doi.org/10.1007/s001800200121>
- O'MALLEY, A. J. and ZASLAVSKY, A. M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *J. Amer. Statist. Assoc.* **103** 1405–1418. [MR2655721](#) <https://doi.org/10.1198/016214508000000724>
- PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. [MR2524001](#) <https://doi.org/10.1198/016214508000000337>
- PFEFFERMANN, D. and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Surv. Methodol.* **16** 217–237.
- PFEFFERMANN, D. and TILLER, R. (2006). Small-area estimation with state-space models subject to benchmark constraints. *J. Amer. Statist. Assoc.* **101** 1387–1397. [MR2307572](#) <https://doi.org/10.1198/016214506000000591>
- POLSON, N. G. and SCOTT, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics* **9** 501–538.
- PROIETTI, T. (2000). Comparing seasonal components for structural time series models. *Int. J. Forecast.* **16** 247–260.
- R CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- RAO, J. N. K. and MOLINA, I. (2015). *Small Area Estimation*, 2nd ed. *Wiley Series in Survey Methodology*. Wiley, Hoboken, NJ. [MR3380626](#) <https://doi.org/10.1002/9781118735855>
- RAO, J. N. K. and YU, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *Canad. J. Statist.* **22** 511–528. [MR1321472](#) <https://doi.org/10.2307/3315407>
- RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields: Theory and Applications. Monographs on Statistics and Applied Probability* **104**. CRC Press, Boca Raton, FL. [MR2130347](#) <https://doi.org/10.1201/9780203492024>
- 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>
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling. Springer Series in Statistics*. Springer, New York. [MR1140409](#) <https://doi.org/10.1007/978-1-4612-4378-6>

- SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. [MR1979380](#) <https://doi.org/10.1111/1467-9868.00353>
- SUGASAWA, S., TAMAE, H. and KUBOKAWA, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scand. J. Stat.* **44** 150–167. [MR3619699](#) <https://doi.org/10.1111/sjos.12246>
- TANG, X., GHOSH, M., HA, N. S. and SEDRANSK, J. (2018). Modeling random effects using global-local shrinkage priors in small area estimation. *J. Amer. Statist. Assoc.* **113** 1476–1489. [MR3902223](#) <https://doi.org/10.1080/01621459.2017.1419135>
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- WATANABE, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11** 3571–3594. [MR2756194](#)
- WOLTER, K. M. (2007). *Introduction to Variance Estimation*, 2nd ed. *Statistics for Social and Behavioral Sciences*. Springer, New York. [MR2288542](#)
- WOOD, S. N. (2017). *Generalized Additive Models: An Introduction with R*. 2nd ed. *Texts in Statistical Science Series*. CRC Press/CRC, Boca Raton, FL. [MR2206355](#)
- YOU, Y. (2008). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Surv. Methodol.* **34** 19–27.
- YOU, Y. and CHAPMAN, B. (2006). Small area estimation using area level models and estimated sampling variances. *Surv. Methodol.* **32** 97–103.
- YOU, Y., RAO, J. and GAMBINO, J. (2003). Model-based unemployment rate estimation for the Canadian labour force survey: A hierarchical Bayes approach. *Surv. Methodol.* **29** 25–32.
- ZHAO, Y., STAUDENMAYER, J., COULL, B. A. and WAND, M. P. (2006). General design Bayesian generalized linear mixed models. *Statist. Sci.* **21** 35–51. [MR2275966](#) <https://doi.org/10.1214/088342306000000015>

HOW MANY REFUGEES AND MIGRANTS DIED TRYING TO REACH EUROPE? JOINT POPULATION SIZE AND TOTAL ESTIMATION

BY ALESSIO FARCOMENI^a

Department of Economics and Finance, University of Rome “Tor Vergata”, ^aalessio.farcomeni@uniroma2.it

We estimate the number of migrants and refugees that died while trying to enter the European Union, during a period of 25 years. Only a subset of attempts with at least one casualty are reported by at least one media source. In order to obtain the estimate, we propose a regression-extrapolation approach, for joint estimation of population size (here, the number of deadly individual or group attempts) and the sum of an accompanying trait (here, the number of deaths) over the population. The trait is measured only for a biased sample of individuals, that are repeatedly observed. Closed-form expressions are derived for the estimator and its standard error. Our findings are that about 40,000 have died from January 1993 to March 2019, during about 5500 attempts to enter the European Union. The number of deaths has been steadily increasing over time, and so has the number of deaths per attempt. About 20% of attempts with at least one casualty have not been recorded by any media source, and slightly less than 10% of deaths have thus been overlooked by media.

REFERENCES

- AMENTA, C., DI BETTA, P. and FERRARA, C. (2021). The migrant crisis in the Mediterranean Sea: Empirical evidence on policy interventions. *Socio-Econ. Plan. Sci.* **78** 101038. <https://doi.org/10.1016/j.seps.2021.101038>
- ANAN, O., BÖHNING, D. and MARUOTTI, A. (2017). Population size estimation and heterogeneity in capture–recapture data: A linear regression estimator based on the Conway–Maxwell–Poisson distribution. *Stat. Methods Appl.* **26** 49–79. MR3610181 <https://doi.org/10.1007/s10260-016-0358-7>
- BALES, K. B., HESKETH, O. and SILVERMAN, B. W. (2015). Modern slavery in the UK: How many victims? *Significance* **12** 16–21.
- BALES, K. B., MURPHY, L. T. and SILVERMAN, B. W. (2020). How many trafficked people are there in Greater New Orleans? Lessons in measurement. *J. Hum. Traffick.* **6** 375–387.
- BARTOLUCCI, F. and LUPPARELLI, M. (2008). Focused information criterion for capture–recapture models for closed populations. *Scand. J. Stat.* **35** 629–649. MR2468866 <https://doi.org/10.1111/j.1467-9469.2008.00604.x>
- BÖHNING, D. (2008). A simple variance formula for population size estimators by conditioning. *Stat. Methodol.* **5** 410–423. MR2528565 <https://doi.org/10.1016/j.stamet.2007.10.001>
- BÖHNING, D. (2016). Ratio plot and ratio regression with applications to social and medical sciences. *Statist. Sci.* **31** 205–218. MR3506100 <https://doi.org/10.1214/16-STS548>
- BÖHNING, D. and FRIEDL, H. (2021). Population size estimation based upon zero-truncated, one-inflated and sparse count data. *Stat. Methods Appl.* **30** 1197–1217. MR4324408 <https://doi.org/10.1007/s10260-021-00556-8>
- BÖHNING, D. and VAN DER HEIJDEN, P. G. M. (2019). The identity of the zero-truncated, one-inflated likelihood and the zero-one-truncated likelihood for general count densities with an application to drink-driving in Britain. *Ann. Appl. Stat.* **13** 1198–1211. MR3963568 <https://doi.org/10.1214/18-AOAS1232>
- BUONFINO, A. (2004). Between unity and plurality: The politicization and securitization of the discourse of immigration in Europe. *New Polit. Sci.* **26** 23–49.
- CELATA, F. and COLETTI, R. (2016). Beyond fortress Europe. Unbounding European normative power and the neighbourhood policy. *Geogr. Compass* **10** 15–24.
- CHAO, A. (1987). Estimating the population size for capture–recapture data with unequal catchability. *Biometrics* **43** 783–791. MR0920467 <https://doi.org/10.2307/2531532>

- CHAO, A. (2001). An overview of closed capture–recapture models. *J. Agric. Biol. Environ. Stat.* **6** 158–175.
- CHAO, A. and COLWELL, R. K. (2017). Thirty years of progeny from Chao’s inequality: Estimating and comparing richness with incidence data and incomplete sampling. *SORT* **41** 3–54. [MR3665906](#)
- COUMANS, A. M., CRUYFF, M. J. L. F., VAN DER HEIJDEN, P. G. M., WOLF, J. and SCHMEETS, H. (2015). Estimating homelessness in the Netherlands using a capture–recapture approach. *Soc. Indic. Res.* **130** 189–212.
- CUSUMANO, E. (2017). Emptying the sea with a spoon? Non-governmental providers of migrants search and rescue in the Mediterranean. *Mar. Policy* **75** 91–98.
- ESPERTI, M. (2019). Rescuing migrants in the central Mediterranean: The emergence of a new civil humanitarianism at the maritime border. *Amer. Behav. Sci.* **64** 436–455.
- FARCOMENI, A. (2018). Fully general Chao and Zelterman estimators with application to a whale shark population. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 217–229. [MR3758763](#) <https://doi.org/10.1111/rssc.12219>
- FARCOMENI, A. (2022). Supplement to “How many refugees and migrants died trying to reach Europe? Joint population size and total estimation.” <https://doi.org/10.1214/21-AOAS1593SUPP>
- FARCOMENI, A. and SCACCIATELLI, D. (2013). Heterogeneity and behavioral response in continuous time capture–recapture, with application to street cannabis use in Italy. *Ann. Appl. Stat.* **7** 2293–2314. [MR3161723](#) <https://doi.org/10.1214/13-AOAS672>
- GODWIN, R. T. (2017). One-inflation and unobserved heterogeneity in population size estimation. *Biom. J.* **59** 79–93. [MR3593722](#) <https://doi.org/10.1002/bimj.201600063>
- HAY, G. and RICHARDSON, C. (2016). Estimating the prevalence of drug use using mark-recapture methods. *Statist. Sci.* **31** 191–204. [MR3506099](#) <https://doi.org/10.1214/16-STS553>
- HERBERS, M. R. (2016). StandUpMigranten: The role of television comedy for the discourse on migration in Germany. *Netw. Knowl.* **9**.
- HUGGINS, R. M., YIP, P. S. F. and STOKLOSA, J. (2016). Nonparametric estimation of the number of drug users in Hong Kong using repeated multiple lists. *Aust. N. Z. J. Stat.* **58** 1–13. [MR3499148](#) <https://doi.org/10.1111/anzs.12149>
- JUNEMANN, A., SCHERER, N. and FROMM, N. (2017). *Fortress Europe? Challenges and Failures of Migration and Asylum Policies*. Springer, Wiesbaden, Germany.
- LACZKO, F., SINGLETON, A., BRIAN, T. and RANGO, M. (2016). Migrant arrivals and deaths in the Mediterranean: What do the data really tell us? *Forced Migr. Rev.* **51** 30–31.
- LAST, T. and SPIJKERBOER, T. (2014). Tracking deaths in the Mediterranean. In *Fatal Journeys. Tracking Lives Lost During Migration* (T. Brian and F. Laczko, eds.) 85–106. International Organization for Migration, Geneva.
- MANRIQUE-VALLIER, D., BALL, P. and SULMONT, D. (2019). Estimating the number of fatal victims of the Peruvian internal armed conflict, 1980–2000: An application of modern multi-list capture–recapture techniques. Preprint. Available at [arXiv:1906.04763](https://arxiv.org/abs/1906.04763).
- MARINO, S. (2016). What are we going to do about them? The centrality of borders in fortress Europe. *Netw. Knowl.* **9**.
- MARINO, S. and DAWES, S. (2016). Introduction to fortress Europe: Media, migration and borders. *Netw. Knowl.* **9**.
- MITCHELL, S., OZONOFF, A., ZASLAVSKY, A. M., HEDT-GAUTHIER, B., LUM, K. and COULL, B. A. (2013). A comparison of marginal and conditional models for capture–recapture data with application to human rights violations data. *Biometrics* **69** 1022–1032. [MR3146797](#) <https://doi.org/10.1111/biom.12089>
- OVERSTALL, A. M., KING, R., BIRD, S. M., HUTCHINSON, S. J. and HAY, G. (2014). Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. *Stat. Med.* **33** 1564–1579. [MR3240769](#) <https://doi.org/10.1002/sim.6047>
- RIVEST, L.-P. and BAILLARGEON, S. (2019). Rcapture: Loglinear models for capture–recapture experiments. R package version 1.4-3.
- ROYLE, J. A. (2009). Analysis of capture–recapture models with individual covariates using data augmentation. *Biometrics* **65** 267–274. [MR2665889](#) <https://doi.org/10.1111/j.1541-0420.2008.01038.x>
- SADINLE, M. (2018). Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. *Ann. Appl. Stat.* **12** 1013–1038. [MR3834293](#) <https://doi.org/10.1214/18-AOAS1178>
- SHILHAV, R. (2017). Beyond fortress Europe: Principles for a humane EU migration policy.
- SILVERMAN, B. W. (2020). Multiple-systems analysis for the quantification of modern slavery: Classical and Bayesian approaches. *J. Roy. Statist. Soc. Ser. A* **183** 691–736. [MR4114463](#)
- SMITH, J. and DAYNES, L. (2016). Borders and migration: An issue of global health importance. *Lancet Glob. Health* **4** 85–86.
- STIERL, M. (2018). A fleet of Mediterranean border humanitarians. *Antipode* **50** 704–724.
- SUKRAZ, B., THAKKINSTIAN, A., OKASCHAROEN, C., AEKPLAKORN, W., SAEJENG, K., BÖHNING, D. and ARUNAKUL, J. (2020). Estimation of the adolescent pregnancy rate in Thailand 2008–2013: An application of capture–recapture method. *BMC Pregnancy Childbirth* **20** 120.

- VAN DER HEIJDEN, P. G. M., BUSTAMI, R., CRUYFF, M. J. L. F., ENGBERSEN, G. and VAN HOUWELINGEN, H. C. (2003). Point and interval estimation of the population size using the truncated Poisson regression model. *Stat. Model.* **3** 305–322. MR2012155 <https://doi.org/10.1191/1471082X03st057oa>
- VIEIRA, I. (2016). The construction of the Mediterranean refugee problem from the Italian digital press (2013–2015): Emergencies in a territory of mobility. *Netw. Knowl.* **9**.

FULL BAYESIAN INFERENCE IN HIDDEN MARKOV MODELS OF PLANT GROWTH

BY GAUTIER VIAUD^{1,a}, YUTING CHEN^{2,c} AND PAUL-HENRY COURNÈDE^{1,b}

¹Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systèmes,
^agautier.viaud@centralesupelec.fr, ^bpaul-henry.cournede@centralesupelec.fr

²Energy Analysis and Environmental Impacts Division, Lawrence Berkeley National Laboratory, ^cyutingchen@lbl.gov

Accurately modeling the growth process of plants in interaction with their environment is important for predicting their biophysical characteristics, referred to as phenotype prediction. Most models are described by discrete dynamic systems in general state-space representation with important domain-specific characteristics: First, plant model parameters have usually clear functional meanings and may be of genetic origins, thus necessitating a precise estimation. Second, critical growth variables, specifically biomass production and dynamic allocation to organs, are hidden variables not accessible to measure. Finally, the difficulty to assess the local plant environment may imply the introduction of process noises in models. Therefore, a precise understanding of the system's behavior requires the joint estimation of functional parameters, hidden states, and noise parameters. In this paper we describe how a full Bayesian method of estimation can accurately estimate all these key model variables using Markov chain Monte Carlo (MCMC) techniques. In the presence of both process and observation noises, it requires to use adequate particle MCMC (PMCMC) algorithms to efficiently sample the hidden states which, consequently, allows for a precise estimation of all noise parameters involved. Thanks to the Bayesian framework, appropriate choices of prior distributions for the noise parameters have enabled analytical posterior distributions and only simple updates are required. Furthermore, this estimation strategy can be easily generalized and adapted to different types of plant growth models, such as organ-scale or compartmental, provided that they are formulated as hidden Markov models. Our estimation method improves on those classically used in plant growth modeling in several aspects: First, by building upon a general probabilistic framework the estimation results allow proper statistical analyses. It is useful in prediction, not only for uncertainty and risk analysis (e.g., for crop yield prediction) but also to analyze the results of experimental trials, for example, to compare genotypes in breeding. Moreover, the care taken in the estimation of hidden variables opens new perspectives in the understanding of inner growth processes, notably the balance and interaction between biomass production and allocation (referred to as source-sink dynamics). Applications of this estimation procedure are demonstrated on the GreenLab model for *Arabidopsis thaliana* and the Log-Normal Allocation and Senescence (LNAS) model for sugar beet, on both synthetic and real data.

REFERENCES

- ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 269–342. [MR2758115](#) <https://doi.org/10.1111/j.1467-9868.2009.00736.x>
- ANDRIEU, C. and ROBERTS, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37** 697–725. [MR2502648](#) <https://doi.org/10.1214/07-AOS574>
- ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* **18** 343–373. [MR2461882](#) <https://doi.org/10.1007/s11222-008-9110-y>

- BAEY, C., DIDIER, A., LEMAIRE, S., MAUPAS, F. and COURNÈDE, P.-H. (2014). Parametrization of five classical plant growth models applied to sugar beet and comparison of their predictive capacity on root yield and total biomass. *Ecol. Model.* **290** 11–20.
- BEAUMONT, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164** 1139–1160.
- CHEN, Y. (2014). Bayesian inference in plant growth models for prediction and uncertainty assessment. Ph.D. thesis, École Centrale Paris.
- CHEN, Y. and COURNÈDE, P. (2014). Data assimilation to reduce uncertainty of crop model prediction with convolution particle filtering. *Ecol. Model.* **290** 165–177.
- COURNÈDE, P.-H., LETORT, V., MATHIEU, A., KANG, M. Z., LEMAIRE, S., TREVEZAS, S., HOULLIER, F. and DE REFFYE, P. (2011). Some parameter estimation issues in functional-structural plant modelling. *Math. Model. Nat. Phenom.* **6** 133–159. [MR2789837 https://doi.org/10.1051/mmnp/20116205](https://doi.org/10.1051/mmnp/20116205)
- COURNÈDE, P.-H., CHEN, Y., WU, Q., BAEY, C. and BAYOL, B. (2013). Development and evaluation of plant growth models: Methodology and implementation in the PYGMALION platform. *Math. Model. Nat. Phenom.* **8** 112–130. [MR3110487 https://doi.org/10.1051/mmnp/20138407](https://doi.org/10.1051/mmnp/20138407)
- DAMAY, N. and LE GOUIS, J. (1993). Radiation use efficiency of sugar beet in northern France. *Eur. J. Agron.* **2** 179–184. [https://doi.org/10.1016/S1161-0301\(14\)80127-5](https://doi.org/10.1016/S1161-0301(14)80127-5)
- DE REFFYE, P., HU, B., KANG, M., LETORT, V. and JAEGER, M. (2021). Two decades of research with the GreenLab model in agronomy. *Ann. Bot.* **127** 281–295.
- DEJONG, T. M., DA SILVA, D., VOS, J. and ESCOBAR-GUTIÉRREZ, A. J. (2011). Using functional–structural plant models to study, understand and integrate plant development and ecophysiology. *Ann. Bot.* **108** 987–989.
- DOUCET, A., DE FREITAS, N. and GORDON, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York.
- DOUCET, A., PITT, M. K., DELIGIANNIDIS, G. and KOHN, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* **102** 295–313. [MR3371005 https://doi.org/10.1093/biomet/asu075](https://doi.org/10.1093/biomet/asu075)
- EVENSEN, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res., Oceans* **99** 10143–10162. <https://doi.org/10.1029/94JC00572>
- FEARNHEAD, P. (2011). MCMC for state-space models. In *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handb. Mod. Stat. Methods 513–529. CRC Press, Boca Raton, FL. [MR2858460](https://doi.org/10.1007/s11222-015-9603-4)
- FEARNHEAD, P. and MELIGKOTSIDOU, L. (2016). Augmentation schemes for particle MCMC. *Stat. Comput.* **26** 1293–1306. [MR3538638 https://doi.org/10.1007/s11222-015-9603-4](https://doi.org/10.1007/s11222-015-9603-4)
- FRANSSON, P., FRANKLIN, O., LINDROOS, O., NILSSON, U. and BRÄNNSTRÖM, Å. (2020). A simulation-based approach to a near-optimal thinning strategy: Allowing harvesting times to be determined for individual trees. *Can. J. For. Res.* **50** 320–331.
- HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. [MR1828504 https://doi.org/10.2307/3318737](https://doi.org/10.2307/3318737)
- HAMMER, G., COOPER, M., TARDIEU, F., WELCH, S., WALSH, B., VAN EEUWIJK, F., CHAPMAN, S. and PODLICH, D. (2006). Models for navigating biological complexity in breeding improved crop plants. *Trends Plant Sci.* **11** 587–593.
- 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](https://doi.org/10.1177/12.280797)
- ILLIAN, J. B., MØLLER, J. and WAAGEPETERSEN, R. P. (2009). Hierarchical spatial point process analysis for a plant community with high biodiversity. *Environ. Ecol. Stat.* **16** 389–405. [MR2749847 https://doi.org/10.1007/s10651-007-0070-8](https://doi.org/10.1007/s10651-007-0070-8)
- JONES, J. W., HOOGENBOOM, G., PORTER, C. H., BOOTE, K. J., BATCHELOR, W. D., HUNT, L., WILKENS, P. W., SINGH, U., GIJSMAN, A. J. et al. (2003). The DSSAT cropping system model. *Eur. J. Agron.* **18** 235–265.
- JONES, J. W., HE, J., BOOTE, K. J., WILKENS, P., PORTER, C. H. and HU, Z. (2015). Estimating DSSAT cropping system cultivar-specific parameters using Bayesian techniques. In *Methods of Introducing System Models into Agricultural Research* 13, 365–393. Wiley, New York.
- JULIER, S. J. and UHLMANN, J. K. (1997). New extension of the Kalman filter to nonlinear systems. *Proc. SPIE* **3068** 182–193. <https://doi.org/10.1117/12.280797>
- KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *Trans. ASME Ser. D. J. Basic Eng.* **82** 35–45. [MR3931993](https://doi.org/10.1117/1.280797)
- KEATING, B. A., CARBERRY, P. S., HAMMER, G. L., PROBERT, M. E., ROBERTSON, M. J., HOLZWORTH, D., HUTH, N. I., HARGREAVES, J. N., MEINKE, H. et al. (2003). An overview of APSIM, a model designed for farming systems simulation. *Eur. J. Agron.* **18** 267–288.

- LEHMANN, N., FINGER, R., KLEIN, T., CALANCA, P. and WALTER, A. (2013). Adapting crop management practices to climate change: Modeling optimal solutions at the field scale. *Agric. Syst.* **117** 55–65.
- LETORT, V., MAHE, P., COURNÈDE, P. H., DE REFFYE, P. and COURTOIS, B. (2008). Quantitative genetics and functional-structural plant growth models: Simulation of quantitative trait loci detection for model parameters and application to potential yield optimization. *Ann. Bot.* **101** 1243–1254.
- LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40. <https://doi.org/10.1093/biomet/81.1.27>
- MARCELIS, L., HEUVELINK, E. and GOUDRIAAN, J. (1998). Modelling of biomass production and yield of horticultural crops: A review. *Sci. Hortic.* **74** 83–111.
- MINGAS, G., BOTTOLO, L. and BOUGANIS, C.-S. (2017). Particle MCMC algorithms and architectures for accelerating inference in state-space models. *Internat. J. Approx. Reason.* **83** 413–433. <https://doi.org/10.1016/j.ijar.2016.10.011>
- NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handb. Mod. Stat. Methods 113–162. CRC Press, Boca Raton, FL. [MR2858447](https://doi.org/10.1016/j.jjar.2016.10.011)
- OVERSTALL, A. M. and WOODS, D. C. (2013). A strategy for Bayesian inference for computationally expensive models with application to the estimation of stem cell properties. *Biometrics* **69** 458–468. [MR3071064](https://doi.org/10.1111/biom.12017)
- PLUCHINOTTA, I., PAGANO, A., GIORDANO, R. and TSOUKIÀS, A. (2018). A system dynamics model for supporting decision-makers in irrigation water management. *J. Environ. Manag.* **223** 815–824. <https://doi.org/10.1016/j.jenvman.2018.06.083>
- PRETZSCH, H., BIBER, P. and ĀURSKÝ, J. (2002). The single tree-based stand simulator SILVA: Construction, application and evaluation. *For. Ecol. Manag.* **162** 3–21.
- QI, R., MA, Y., HU, B., DE REFFYE, P. and COURNÈDE, P. H. (2010). Optimization of source-sink dynamics in plant growth for ideotype breeding: A case study on maize. *Comput. Electron. Agric.* **71** 96–105.
- QUILLOT-TURION, B., OULD-SIDI, M.-M., KADRANI, A., HILGERT, N., GÉNARD, M. and LESCOURET, F. (2012). Optimization of parameters of the Virtual Fruit model to design peach genotype for sustainable production systems. *Eur. J. Agron.* **42** 34–48.
- RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77** 257–286.
- ROBERTS, G. O. and SAHU, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. Roy. Statist. Soc. Ser. B* **59** 291–317. [MR1440584](https://doi.org/10.1111/1467-9868.00070)
- SHERLOCK, C., THIERY, A. H. and LEE, A. (2017). Pseudo-marginal Metropolis–Hastings sampling using averages of unbiased estimators. *Biometrika* **104** 727–734. [MR3694593](https://doi.org/10.1093/biomet/asx031)
- SMITH, M. R., RAO, I. M. and MERCHANT, A. (2018). Source-sink relationships in crop plants and their influence on yield development and nutritional quality. *Front. Plant Sci.* **9** 1889. <https://doi.org/10.3389/fpls.2018.01889>
- TISNÉ, S., SERRAND, Y., BACH, L., GILBAULT, E., BEN AMEUR, R., BALASSE, H., VOISIN, R., BOUCHEZ, D., GUERCHE, P. et al. (2013). Phenoscope: An automated large-scale phenotyping platform offering high spatial homogeneity. *Plant J.* **74** 534–544.
- VIAUD, G. (2018). Statistical methods for the genotypic differentiation of plants using growth models. Ph.D. thesis, Univ. Paris-Saclay.
- VIAUD, G., CHEN, Y. and COURNÈDE, P.-H. (2022). Supplement to “Full Bayesian inference in hidden Markov models of plant growth.” <https://doi.org/10.1214/21-AOAS1594SUPP>
- VIAUD, G., LOUDET, O. and COURNÈDE, P.-H. (2017). Leaf segmentation and tracking in *Arabidopsis thaliana* combined to an organ-scale plant model for genotypic differentiation. *Front. Plant Sci.* **7** 2057.
- WALLACH, D. (2006). Evaluating crop models. In *Working with Dynamic Crop Models: Evaluation, Analysis, Parameterization, and Applications* 2, 11–54. Elsevier, Amsterdam.
- WHITE, A. C., ROGERS, A., REES, M. and OSBORNE, C. P. (2016). How can we make plants grow faster? A source-sink perspective on growth rate. *J. Exp. Bot.* **67** 31–45. <https://doi.org/10.1093/jxb/erv447>
- WILHELM, W. and MCMASTER, G. S. (1995). Importance of the phyllochron in studying development and growth in grasses. *Crop Sci.* **35** 1–3.
- WRIGHT, I. J., REICH, P. B., WESTOBY, M., ACKERLY, D. D., BARUCH, Z., BONGERS, F., CAVENDER-BARES, J., CHAPIN, T., CORNELISSEN, J. H. et al. (2004). The worldwide leaf economics spectrum. *Nature* **428** 821–827.
- WU, L., LE DIMET, F.-X., DE REFFYE, P., HU, B.-G., COURNÈDE, P.-H. and KANG, M.-Z. (2012). An optimal control methodology for plant growth—Case study of a water supply problem of sunflower. *Math. Comput. Simulation* **82** 909–923. [MR2892130](https://doi.org/10.1016/j.matcom.2011.12.007)
- YIN, X., STRUIK, P. C. and KROPFF, M. J. (2004). Role of crop physiology in predicting gene-to-phenotype relationships. *Trends Plant Sci.* **9** 426–432. <https://doi.org/10.1016/j.tplants.2004.07.007>

EXTENDED STOCHASTIC BLOCK MODELS WITH APPLICATION TO CRIMINAL NETWORKS

BY SIRIO LEGRAMANTI^{1,a}, TOMMASO RIGON^{2,c}, DANIELE DURANTE^{1,b} AND DAVID B. DUNSON^{3,d}

¹Department Decision Sciences and Institute for Data Science and Analytics, Bocconi University,
^asirio.legramanti@unibocconi.it, ^bdaniele.durante@unibocconi.it

²Department of Economics, Management and Statistics, University of Milano-Bicocca, ^ctommaso.rigon@unimib.it

³Department of Statistical Science, Duke University, ^ddunson@duke.edu

Reliably learning group structures among nodes in network data is challenging in several applications. We are particularly motivated by studying covert networks that encode relationships among criminals. These data are subject to measurement errors, and exhibit a complex combination of an unknown number of core-periphery, assortative and disassortative structures that may unveil key architectures of the criminal organization. The coexistence of these noisy block patterns limits the reliability of routinely-used community detection algorithms, and requires extensions of model-based solutions to realistically characterize the node partition process, incorporate information from node attributes, and provide improved strategies for estimation and uncertainty quantification. To cover these gaps, we develop a new class of extended stochastic block models (ESBM) that infer groups of nodes having common connectivity patterns via Gibbs-type priors on the partition process. This choice encompasses many realistic priors for criminal networks, covering solutions with fixed, random and infinite number of possible groups, and facilitates the inclusion of node attributes in a principled manner. Among the new alternatives in our class, we focus on the Gnedin process as a realistic prior that allows the number of groups to be finite, random and subject to a reinforcement process coherent with criminal networks. A collapsed Gibbs sampler is proposed for the whole ESBM class, and refined strategies for estimation, prediction, uncertainty quantification and model selection are outlined. The ESBM performance is illustrated in realistic simulations and in an application to an Italian mafia network, where we unveil key complex block structures, mostly hidden from state-of-the-art alternatives.

REFERENCES

- ABBE, E. (2017). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18** 1–86. [MR3827065](#)
- AGRESTE, S., CATANESE, S., DE MEO, P., FERRARA, E. and FIUMARA, G. (2016). Network structure and resilience of mafia syndicates. *Inform. Sci.* **351** 30–47.
- AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic block-models. *J. Mach. Learn. Res.* **9** 1981–2014.
- AMINI, A., CHEN, A., BICKEL, P. J. and LEVINA, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41** 2097–2122. [MR3127859](#) <https://doi.org/10.1214/13-AOS1138>
- ATHREYA, A., FISHKIND, D. E., TANG, M., PRIEBE, C. E., PARK, Y., VOGELSTEIN, J. T., LEVIN, K., LYZINSKI, V., QIN, Y. and SUSSMAN, D. L. (2017). Statistical inference on random dot product graphs: A survey. *J. Mach. Learn. Res.* **18** 1–92. [MR3827114](#)
- BICKEL, P. J., CHOI, D., CHANG, X. and ZHANG, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.* **41** 1922–1943. [MR3127853](#) <https://doi.org/10.1214/13-AOS1124>
- BINKIEWICZ, N., VOGELSTEIN, J. T. and ROHE, K. (2017). Covariate-assisted spectral clustering. *Biometrika* **104** 361–377. [MR3698259](#) <https://doi.org/10.1093/biomet/asx008>

- BLONDEL, V. D., GUILLAUME, J. L., LAMBIOTTE, R. and LEFEBVRE, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* **10** P10008.
- CALDERONI, F., BRUNETTO, D. and PICCARDI, C. (2017). Communities in criminal networks: A case study. *Soc. Netw.* **48** 116–125.
- CALDERONI, F. and PICCARDI, C. (2014). Uncovering the structure of criminal organizations by community analysis: The Infinito network. In *2014 Tenth International Conference on Signal–Image Technology and Internet–Based Systems* 301–308. IEEE, Marrakech, Morocco.
- CAMPANA, P. (2016). Explaining criminal networks: Strategies and potential pitfalls. *Methodol. Innov.* **9** 1–10.
- CAMPANA, P. and VARESE, F. (2022). Studying organized crime networks: Data sources, boundaries and the limits of structural measures. *Soc. Netw.* **69** 149–159.
- CARLEY, K. M., LEE, J.-S. and KRACKHARDT, D. (2002). Destabilizing networks. *Connections* **24** 79–92.
- CATINO, M. (2014). How do mafias organize? Conflict and violence in three mafia organizations. *Eur. J. Sociol.* **55** 177–220.
- CAVALLARO, L., FICARA, A., DE MEO, P., FIUMARA, G., CATANESE, S., BAGDASAR, O., SONG, W. and LIOTTA, A. (2020). Disrupting resilient criminal networks through data analysis: The case of Sicilian Mafia. *PLoS ONE* **15** 1–22.
- CHEN, K. and LEI, J. (2018). Network cross-validation for determining the number of communities in network data. *J. Amer. Statist. Assoc.* **113** 241–251. [MR3803461 https://doi.org/10.1080/01621459.2016.1246365](https://doi.org/10.1080/01621459.2016.1246365)
- CÔME, E. and LATOUCHE, P. (2015). Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Stat. Model.* **15** 564–589. [MR3441229 https://doi.org/10.1177/1471082X15577017](https://doi.org/10.1177/1471082X15577017)
- CÔME, E., JOUVIN, N., LATOUCHE, P. and BOUVEYRON, C. (2021). Hierarchical clustering with discrete latent variable models and the integrated classification likelihood. *Adv. Data Anal. Classif.* **15** 957–986. [MR4333226 https://doi.org/10.1007/s11634-021-00440-z](https://doi.org/10.1007/s11634-021-00440-z)
- DE BLASI, P., LIJOI, A. and PRÜNSTER, I. (2013). An asymptotic analysis of a class of discrete nonparametric priors. *Statist. Sinica* **23** 1299–1321. [MR3114715](https://doi.org/10.1007/s114715)
- DE BLASI, P., FAVARO, S., LIJOI, A., MENA, R. H., PRÜNSTER, I. and RUGGIERO, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 212–229.
- DIVIÁK, T. (2022). Key aspects of covert networks data collection: Problems, challenges, and opportunities. *Soc. Netw.* **69** 160–169.
- FAUST, K. and TITA, G. E. (2019). Social networks and crime: Pitfalls and promises for advancing the field. *Annu. Rev. Criminol.* **2** 99–122.
- FERRARA, E., DE MEO, P., CATANESE, S. and FIUMARA, G. (2014). Detecting criminal organizations in mobile phone networks. *Expert Syst. Appl.* **41** 5733–5750.
- FORTUNATO, S. and HRIC, D. (2016). Community detection in networks: A user guide. *Phys. Rep.* **659** 1–44. [MR3566093 https://doi.org/10.1016/j.physrep.2016.09.002](https://doi.org/10.1016/j.physrep.2016.09.002)
- FOSDICK, B. K., MCCORMICK, T. H., MURPHY, T. B., NG, T. L. J. and WESTLING, T. (2019). Multiresolution network models. *J. Comput. Graph. Statist.* **28** 185–196. [MR3939381 https://doi.org/10.1080/10618600.2018.1505633](https://doi.org/10.1080/10618600.2018.1505633)
- FRUCHTERMAN, T. M. and REINGOLD, E. M. (1991). Graph drawing by force-directed placement. *Softw. Pract. Exp.* **21** 1129–1164.
- GELMAN, A., HWANG, J. and VEHTARI, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24** 997–1016. [MR3253850 https://doi.org/10.1007/s11222-013-9416-2](https://doi.org/10.1007/s11222-013-9416-2)
- GENG, J., BHATTACHARYA, A. and PATI, D. (2019). Probabilistic community detection with unknown number of communities. *J. Amer. Statist. Assoc.* **114** 893–905. [MR3963189 https://doi.org/10.1080/01621459.2018.1458618](https://doi.org/10.1080/01621459.2018.1458618)
- GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826. [MR1908073 https://doi.org/10.1073/pnas.122653799](https://doi.org/10.1073/pnas.122653799)
- GNEDIN, A. (2010). A species sampling model with finitely many types. *Electron. Commun. Probab.* **15** 79–88. [MR2606505 https://doi.org/10.1214/ECP.v15-1532](https://doi.org/10.1214/ECP.v15-1532)
- GNEDIN, A. and PITMAN, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. (POMI) S.-Peterburg.* **325** 83–102. [MR2160320 https://doi.org/10.1007/s10958-006-0335-z](https://doi.org/10.1007/s10958-006-0335-z)
- GORMLEY, I. C. and MURPHY, T. B. (2010). A mixture of experts latent position cluster model for social network data. *Stat. Methodol.* **7** 385–405. [MR2643609 https://doi.org/10.1016/j.stamet.2010.01.002](https://doi.org/10.1016/j.stamet.2010.01.002)
- GRASSI, R., CALDERONI, F., BIANCHI, M. and TORRIERO, A. (2019). Betweenness to assess leaders in criminal networks: New evidence using the dual projection approach. *Soc. Netw.* **56** 23–32.
- HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354. [MR2364300 https://doi.org/10.1111/j.1467-985X.2007.00471.x](https://doi.org/10.1111/j.1467-985X.2007.00471.x)

- HARTIGAN, J. A. (1990). Partition models. *Comm. Statist. Theory and Methods* **19** 2745–2756. [MR1088047](#) <https://doi.org/10.1080/03610929008830345>
- HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. [MR0718088](#) [https://doi.org/10.1016/0378-8733\(83\)90021-7](https://doi.org/10.1016/0378-8733(83)90021-7)
- KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107, 10. [MR2788206](#) <https://doi.org/10.1103/PhysRevE.83.016107>
- KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. [MR3363402](#) <https://doi.org/10.1080/01621459.1995.10476572>
- KEMP, C., TENENBAUM, J. B., GRIFFITHS, T. L., YAMADA, T. and UEDA, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence* 381–388.
- KIM, D., HUGHES, M. and SUDDERTH, E. (2012). The nonparametric metadata dependent relational model. In *ICML'12: Proceedings of the 29th International Conference on Machine Learning* 1411–1418. IEEE, Edinburgh, UK.
- KREBS, V. E. (2002). Mapping networks of terrorist cells. *Connections* **24** 43–52.
- LE, V. (2012). Organised crime typologies: Structure, activities and conditions. *Int. J. Criminol. Sociol.* **1** 121–131.
- LE, C. M. and LEVINA, E. (2015). Estimating the number of communities in networks by spectral methods. Preprint. Available at [arXiv:1507.00827](#).
- LEE, C. and WILKINSON, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Appl. Netw. Sci.* **4** 1–50.
- LEGAMANTI, S., RIGON, T. and DURANTE, D. (2020). Bayesian testing for exogenous partition structures in stochastic block models. *Sankhya A*. In press.
- LEGAMANTI, S., RIGON, T., DURANTE, D. and DUNSON, D. B. (2022). Supplement to “Extended stochastic block models with application to criminal networks.” <https://doi.org/10.1214/21-AOAS1595SUPP>
- LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. [MR3285605](#) <https://doi.org/10.1214/14-AOS1274>
- LENK, P. (2009). Simulation pseudo-bias correction to the harmonic mean estimator of integrated likelihoods. *J. Comput. Graph. Statist.* **18** 941–960. [MR2750446](#) <https://doi.org/10.1198/jcgs.2009.08022>
- LI, T., LEVINA, E. and ZHU, J. (2020). Network cross-validation by edge sampling. *Biometrika* **107** 257–276. [MR4108931](#) <https://doi.org/10.1093/biomet/asaa006>
- LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007a). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 715–740. [MR2370077](#) <https://doi.org/10.1111/j.1467-9868.2007.00609.x>
- LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007b). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94** 769–786. [MR2416792](#) <https://doi.org/10.1093/biomet/asm061>
- LIJOI, A., PRÜNSTER, I. and WALKER, S. G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.* **18** 1519–1547. [MR2434179](#) <https://doi.org/10.1214/07-AAP495>
- LIU, F., CHOI, D., XIE, L. and ROEDER, K. (2018). Global spectral clustering in dynamic networks. *Proc. Natl. Acad. Sci. USA* **115** 927–932. [MR3763702](#) <https://doi.org/10.1073/pnas.1718449115>
- MAGALINGAM, P., DAVIS, S. and RAO, A. (2015). Using shortest path to discover criminal community. *Digit. Investig.* **15** 1–17.
- MALM, A. and BICHLER, G. (2011). Networks of collaborating criminals: Assessing the structural vulnerability of drug markets. *J. Res. Crime Delinq.* **48** 271–297.
- MEILÄ, M. (2007). Comparing clusterings—an information based distance. *J. Multivariate Anal.* **98** 873–895. [MR2325412](#) <https://doi.org/10.1016/j.jmva.2006.11.013>
- MILLER, J. W. and HARRISON, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *J. Mach. Learn. Res.* **15** 3333–3370. [MR3277163](#)
- MILLER, J. W. and HARRISON, M. T. (2018). Mixture models with a prior on the number of components. *J. Amer. Statist. Assoc.* **113** 340–356. [MR3803469](#) <https://doi.org/10.1080/01621459.2016.1255636>
- MORSELLI, C. (2009). Hells Angels in springtime. *Trends Organ. Crime* **12** 145–158.
- MORSELLI, C., GIGUÈRE, C. and PETIT, K. (2007). The efficiency/security trade-off in criminal networks. *Soc. Netw.* **29** 143–153.
- MÜLLER, P., QUINTANA, F. and ROSNER, G. L. (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist.* **20** 260–278. [MR2816548](#) <https://doi.org/10.1198/jcgs.2011.09066>
- NEWMAN, M. E. J. and GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* **69** 026113.
- NEWMAN, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577–8582.

- NEWMAN, M. E. J. and CLAUSET, A. (2016). Structure and inference in annotated networks. *Nat. Commun.* **7** 1–11.
- NOROOZI, M. and PENSKEY, M. (2020). Statistical inference in heterogeneous block model. Preprint. Available at [arXiv:2002.02610](https://arxiv.org/abs/2002.02610).
- NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96** 1077–1087. [MR1947255 https://doi.org/10.1198/016214501753208735](https://doi.org/10.1198/016214501753208735)
- OLHEDE, S. C. and WOLFE, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proc. Natl. Acad. Sci. USA* **111** 14722–14727.
- PAJOR, A. (2017). Estimating the marginal likelihood using the arithmetic mean identity. *Bayesian Anal.* **12** 261–287. [MR3620130 https://doi.org/10.1214/16-BA1001](https://doi.org/10.1214/16-BA1001)
- PAOLI, L. (2007). Mafia and organised crime in Italy: The unacknowledged successes of law enforcement. *West Eur. Polit.* **30** 854–880.
- PARK, J.-H. and DUNSON, D. B. (2010). Bayesian generalized product partition model. *Statist. Sinica* **20** 1203–1226. [MR2730180](https://doi.org/10.1214/10-BA1001)
- QUINTANA, F. A. and IGLESIAS, P. L. (2003). Bayesian clustering and product partition models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 557–574. [MR1983764 https://doi.org/10.1111/1467-9868.00402](https://doi.org/10.1111/1467-9868.00402)
- RAFTERY, A. E., NEWTON, M. A., SATAGOPAN, J. M. and KRIVITSKY, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In *Bayesian Statistics 8*. 1–45. Oxford Univ. Press, Oxford. [MR2433201](https://doi.org/10.1093/ba/bes001)
- RANCIATI, S., VINCIOTTI, V. and WIT, E. C. (2020). Identifying overlapping terrorist cells from the Noordin Top actor-event network. *Ann. Appl. Stat.* **14** 1516–1534. [MR4152144 https://doi.org/10.1214/20-AOAS1358](https://doi.org/10.1214/20-AOAS1358)
- RASTELLI, R., LATOUCHE, P. and FRIEL, N. (2018). Choosing the number of groups in a latent stochastic blockmodel for dynamic networks. *Netw. Sci.* **6** 469–493.
- ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic block-model. *Ann. Statist.* **39** 1878–1915. [MR2893856 https://doi.org/10.1214/11-AOS887](https://doi.org/10.1214/11-AOS887)
- SALDAÑA, D. F., YU, Y. and FENG, Y. (2017). How many communities are there? *J. Comput. Graph. Statist.* **26** 171–181. [MR3610418 https://doi.org/10.1080/10618600.2015.1096790](https://doi.org/10.1080/10618600.2015.1096790)
- SANGKARAN, T., ABDULLAH, A. and JHANJHI, N. (2020). Criminal community detection based on isomorphic subgraph analytics. *Open Comput. Sci.* **10** 164–174.
- SARKAR, P. and BICKEL, P. J. (2015). Role of normalization in spectral clustering for stochastic blockmodels. *Ann. Statist.* **43** 962–990. [MR3346694 https://doi.org/10.1214/14-AOS1285](https://doi.org/10.1214/14-AOS1285)
- SCHMIDT, M. N. and MORUP, M. (2013). Nonparametric Bayesian modeling of complex networks: An introduction. *IEEE Signal Process. Mag.* **30** 110–128.
- SENGUPTA, S. and CHEN, Y. (2018). A block model for node popularity in networks with community structure. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 365–386. [MR3763696 https://doi.org/10.1111/rssb.12245](https://doi.org/10.1111/rssb.12245)
- SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. [MR1979380 https://doi.org/10.1111/1467-9868.00353](https://doi.org/10.1111/1467-9868.00353)
- STANLEY, N., BONACCI, T., KWITT, R., NIETHAMMER, M. and MUCHA, P. J. (2019). Stochastic block models with multiple continuous attributes. *Appl. Netw. Sci.* **4** 1–22.
- SUSSMAN, D. L., TANG, M., FISHKIND, D. and PRIEBE, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* **107** 1119–1128. [MR3010899 https://doi.org/10.1080/01621459.2012.699795](https://doi.org/10.1080/01621459.2012.699795)
- TALLBERG, C. (2004). A Bayesian approach to modeling stochastic blockstructures with covariates. *J. Math. Sociol.* **29** 1–23.
- VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. [MR2409803 https://doi.org/10.1007/s11222-007-9033-z](https://doi.org/10.1007/s11222-007-9033-z)
- WADE, S. and GHAHRAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Anal.* **13** 559–626. [MR3807860 https://doi.org/10.1214/17-BA1073](https://doi.org/10.1214/17-BA1073)
- WANG, Y. X. R. and BICKEL, P. J. (2017). Likelihood-based model selection for stochastic block models. *Ann. Statist.* **45** 500–528. [MR3650391 https://doi.org/10.1214/16-AOS1457](https://doi.org/10.1214/16-AOS1457)
- WANG, Y.-B., CHEN, M.-H., KUO, L. and LEWIS, P. O. (2018). A new Monte Carlo method for estimating marginal likelihoods. *Bayesian Anal.* **13** 311–333. [MR3780425 https://doi.org/10.1214/17-BA1049](https://doi.org/10.1214/17-BA1049)
- WATANABE, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11** 3571–3594. [MR2756194](https://doi.org/10.1214/10-AOS1285)
- WATANABE, S. (2013). A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14** 867–897. [MR3049492](https://doi.org/10.1214/10-AOS1285)
- WHITE, A. and MURPHY, T. B. (2016). Mixed-membership of experts stochastic blockmodel. *Netw. Sci.* **4** 48–80.

- XU, Z., KE, Y., WANG, Y., CHENG, H. and CHENG, J. (2012). A model-based approach to attributed graph clustering. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* 505–516.
- YANG, J., MCAULEY, J. and LESKOVEC, J. (2013). Community detection in networks with node attributes. In *2013 IEEE 13th International Conference on Data Mining* 1151–1156.
- ZHANG, Y., LEVINA, E. and ZHU, J. (2016). Community detection in networks with node features. *Electron. J. Stat.* **10** 3153–3178. [MR3571965](#) <https://doi.org/10.1214/16-EJS1206>
- ZHAO, H., DU, L. and BUNTINE, W. (2017). Leveraging node attributes for incomplete relational data. In *International Conference on Machine Learning* 4072–4081.
- ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. [MR3059083](#) <https://doi.org/10.1214/12-AOS1036>
- ZHOU, Z. and AMINI, A. (2019). Analysis of spectral clustering algorithms for community detection: The general bipartite setting. *J. Mach. Learn. Res.* **20** 1–47. [MR3948087](#)

TWO-SAMPLE TESTS FOR MULTIVARIATE REPEATED MEASUREMENTS OF HISTOGRAM OBJECTS WITH APPLICATIONS TO WEARABLE DEVICE DATA

BY JINGRU ZHANG^{1,a}, KATHLEEN R. MERIKANGAS^{2,d}, HONGZHE LI^{1,b} AND
HAOCHANG SHOU^{1,c}

¹*Division of Biostatistics, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, ^ajingru.zhang@pennmedicine.upenn.edu, ^bhongzhe@pennmedicine.upenn.edu, ^chshou@pennmedicine.upenn.edu*

²*Genetic Epidemiology Research Branch, National Institute of Mental Health, National Institutes of Health, ^dmerikank@mail.nih.gov*

Repeated observations have become increasingly common in biomedical research and longitudinal studies. For instance, wearable sensor devices are deployed to continuously track physiological and biological signals from each individual over multiple days. It remains of great interest to appropriately evaluate how the daily distribution of biosignals might differ across disease groups and demographics. Hence, these data could be formulated as multivariate complex object data, such as probability densities, histograms, and observations on a tree. Traditional statistical methods would often fail to apply, as they are sampled from an arbitrary non-Euclidean metric space. In this paper we propose novel, nonparametric, graph-based two-sample tests for object data with the same structure of repeated measures. We treat the repeatedly measured object data as multivariate object data, which requires the same number of repeated observations per individual but eliminates any assumptions on the errors of the repeated observations. A set of test statistics are proposed to capture various possible alternatives. We derive their asymptotic null distributions under the permutation null. These tests exhibit substantial power improvements over the existing methods while controlling the type I errors under finite samples as shown through simulation studies. The proposed tests are demonstrated to provide additional insights on the location, inter- and intra-individual variability of the daily physical activity distributions in a sample of studies for mood disorders.

REFERENCES

- BANDA, J. A., HAYDEL, K. F., DAVILA, T., DESAI, M., BRYSON, S., HASKELL, W. L., MATHESON, D. and ROBINSON, T. N. (2016). Effects of varying epoch lengths, wear time algorithms, and activity cut-points on estimates of child sedentary behavior and physical activity from accelerometer data. *PLoS ONE* **11** e0150534.
- BURTON, C., MCKINSTRY, B., TĂTAR, A. S., SERRANO-BLANCO, A., PAGLIARI, C. and WOLTERS, M. (2013). Activity monitoring in patients with depression: A systematic review. *J. Affective Disorders* **145** 21–28.
- CHAPMAN, J. J., ROBERTS, J. A., NGUYEN, V. T. and BREAKSPEAR, M. (2017). Quantification of free-living activity patterns using accelerometry in adults with mental illness. *Sci. Rep.* **7** 43174.
- CHEN, H., CHEN, X. and SU, Y. (2018). A weighted edge-count two-sample test for multivariate and object data. *J. Amer. Statist. Assoc.* **113** 1146–1155. [MR3862346 https://doi.org/10.1080/01621459.2017.1307757](https://doi.org/10.1080/01621459.2017.1307757)
- CHEN, H. and FRIEDMAN, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *J. Amer. Statist. Assoc.* **112** 397–409. [MR3646580 https://doi.org/10.1080/01621459.2016.1147356](https://doi.org/10.1080/01621459.2016.1147356)
- CRESCENZO, F. D., ECONOMOU, A., SHARPLEY, A. L., GORMEZ, A. and QUESTED, D. J. (2017). Actigraphic features of bipolar disorder: A systematic review and meta-analysis. *Sleep Med. Rev.* **33** 58–69. <https://doi.org/10.1016/j.smrv.2016.05.003>

- DAWSON, J. D. and LAGAKOS, S. W. (1993). Size and power of two-sample tests of repeated measures data. *Biometrics* **49** 1022–1032. [MR1257692 https://doi.org/10.2307/2532244](https://doi.org/10.2307/2532244)
- DOHERTY, A., JACKSON, D., HAMMERLA, N., PLÖTZ, T., OLIVIER, P., GRANAT, M. H., WHITE, T., VAN HEES, V. T., TRENNELL, M. I. et al. (2017). Large scale population assessment of physical activity using wrist worn accelerometers: The UK biobank study. *PLoS ONE* **12** e0169649. <https://doi.org/10.1371/journal.pone.0169649>
- DUBEY, P. and MÜLLER, H.-G. (2019). Fréchet analysis of variance for random objects. *Biometrika* **106** 803–821. [MR4031200 https://doi.org/10.1093/biomet/asz052](https://doi.org/10.1093/biomet/asz052)
- FAURHOLT-JEPSEN, M., BRAGE, S., VINBERG, M., CHRISTENSEN, E. M., KNORR, U., JENSEN, H. M. and KESSING, L. V. (2012). Differences in psychomotor activity in patients suffering from unipolar and bipolar affective disorder in the remitted or mild/moderate depressive state. *J. Affective Disorders* **141** 457–463. <https://doi.org/10.1016/j.jad.2012.02.020>
- FRIEDMAN, J. H. and RAFSKY, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7** 697–717. [MR0532236](https://doi.org/10.2307/2346178)
- GRETTON, A., BORWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. [MR2913716](https://doi.org/10.26434/chemrxiv-2012-02-020)
- HENZE, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.* **16** 772–783. [MR0947577 https://doi.org/10.1214/aos/1176350835](https://doi.org/10.1214/aos/1176350835)
- INDIC, P., SALVATORE, P., MAGGINI, C., GHIDINI, S., FERRARO, G., BALDESSARINI, R. J. and MURRAY, G. (2011). Scaling behavior of human locomotor activity amplitude: Association with bipolar disorder. *PLoS ONE* **6** e20650. <https://doi.org/10.1371/journal.pone.0020650>
- KEADLE, S. K., SHIROMA, E. J., FREEDSON, P. S. and LEE, I. M. (2014). Impact of accelerometer data processing decisions on the sample size, wear time and physical activity level of a large cohort study. *BMC Public Health* **14** 1210. <https://doi.org/10.1186/1471-2458-14-1210>
- KRANE-GARTISER, K., HENRIKSEN, T. E. G., MORKEN, G., VAALER, A. and FASMER, O. B. (2014). Actigraphic assessment of motor activity in acutely admitted inpatients with bipolar disorder. *PLoS ONE* **9** e89574. <https://doi.org/10.1371/journal.pone.0089574>
- LEEGER-ASCHMANN, C. S., SCHMUTZ, E. A., ZYSSET, A. E., KAKEBEEKE, T. H., MESSERLI-BÜRGY, N., STÜLB, K., ARHAB, A., MEYER, A. H., MUNSCH, S. et al. (2019). Accelerometer-derived physical activity estimation in preschoolers—comparison of cut-point sets incorporating the vector magnitude vs the vertical axis. *BMC Public Health* **19** 513. <https://doi.org/10.1186/s12916-019-1313-1>
- MERIKANGAS, K. R., CUI, L., HEATON, L., NAKAMURA, E., ROCA, C., DING, J., QIN, H., GUO, W., SHUGART, Y. Y. et al. (2014). Independence of familial transmission of mania and depression: Results of the NIMH family study of affective spectrum disorders. *Mol. Psychiatry* **19** 214–9. <https://doi.org/10.1038/mp.2013.116>
- MERIKANGAS, K. R., SWENDSEN, J., HICKIE, I. B., CUI, L., SHOU, H., MERIKANGAS, A. K., ZHANG, J., LAMERS, F., CRAINICEANU, C. et al. (2019). Real-time mobile monitoring of the dynamic associations among motor activity, energy, mood, and sleep in adults with bipolar disorder. *JAMA Psychiatr.* **76** 190–198. <https://doi.org/10.1001/jama.2019.130247>
- MURRAY, G., GOTTLIEB, J., HIDALGO, M. P., ETAIN, B., RITTER, P., SKENE, D. J., GARBAZZA, C., BULLOCK, B., MERIKANGAS, K. et al. (2020). Measuring circadian function in bipolar disorders: Empirical and conceptual review of physiological, actigraphic, and self-report approaches. *Bipolar Disorders*. <https://doi.org/10.1111/bdi.12963>
- PAGANI, L., CLAIR, P. A. S., TESHIBA, T. M., SERVICE, S. K., FEARS, S. C., ARAYA, C., ARAYA, X., BEJARANO, J., RAMIREZ, M. et al. (2016). Genetic contributions to circadian activity rhythm and sleep pattern phenotypes in pedigrees segregating for severe bipolar disorder. *Proc. Natl. Acad. Sci. USA* **113** E754–E761. <https://doi.org/10.1073/pnas.1513525113>
- PETERSEN, A. and MÜLLER, H.-G. (2019). Wasserstein covariance for multiple random densities. *Biometrika* **106** 339–351. [MR3949307 https://doi.org/10.1093/biomet/asz005](https://doi.org/10.1093/biomet/asz005)
- ROBILLARD, R., HERMENS, D. F., NAISMITH, S. L., WHITE, D., ROGERS, N. L., IP, T. K. C., MULLIN, S. J., ALVARES, G. A., GUASTELLA, A. J. et al. (2015). Ambulatory sleep-wake patterns and variability in young people with emerging mental disorders. *J. Psychiatry Neurosci.* **40** 28–37. <https://doi.org/10.1503/jpn.130247>
- ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 515–530. [MR2168202 https://doi.org/10.1111/j.1467-9868.2005.00513.x](https://doi.org/10.1111/j.1467-9868.2005.00513.x)
- SCHILLING, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 799–806. [MR0860514](https://doi.org/10.2307/2287014)
- SCHOTT, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Comput. Statist. Data Anal.* **51** 6535–6542. [MR2408613 https://doi.org/10.1016/j.csda.2007.03.004](https://doi.org/10.1016/j.csda.2007.03.004)

- SCHRACK, J. A., ZIPUNNIKOV, V., GOLDSMITH, J., BAI, J., SIMONSICK, E. M., CRAINICEANU, C. and FERRUCCI, L. (2014). Assessing the “physical cliff”: Detailed quantification of age-related differences in daily patterns of physical activity. *J. Gerontol., Ser. A, Biol. Sci. Med. Sci.* **69** 973–9. <https://doi.org/10.1093/gerona/glt199>
- SCHRACK, J. A., COOPER, R., KOSTER, A., SHIROMA, E. J., MURABITO, J. M., REJESKI, W. J., FERRUCCI, L. and HARRIS, T. B. (2016). Assessing daily physical activity in older adults: Unraveling the complexity of monitors, measures, and methods. *J. Gerontol., Ser. A, Biol. Sci. Med. Sci.* **71** 1039–1048. <https://doi.org/10.1093/gerona/glw026>
- SCOTT, J., MURRAY, G., HENRY, C., MORKEN, G., SCOTT, E., ANGST, J., MERIKANGAS, K. R. and HICKIE, I. B. (2017). Activation in bipolar disorders: A systematic review. *JAMA Psychiatr.* **74** 189–196. <https://doi.org/10.1001/jamapsychiatry.2016.3459>
- SHOU, H., CUI, L., HICKIE, I., LAMEIRA, D., LAMERS, F., ZHANG, J., CRAINICEANU, C., ZIPUNNIKOV, V. and MERIKANGAS, K. R. (2017). Dysregulation of objectively assessed 24-hour motor activity patterns as a potential marker for bipolar I disorder: Results of a community-based family study. *Translational Psychiatry* **7** e1211. <https://doi.org/10.1038/tp.2017.136>
- VARMA, V. R., DEY, D., LEROUX, A., DI, J., URBANEK, J., XIAO, L. and ZIPUNNIKOV, V. (2017). Re-evaluating the effect of age on physical activity over the lifespan. *Prev. Med.* **101** 102–108. <https://doi.org/10.1016/j.ypmed.2017.05.030>
- VARMA, V. R., DEY, D., LEROUX, A., DI, J., URBANEK, J., XIAO, L. and ZIPUNNIKOV, V. (2018). Total volume of physical activity: TAC, TLAC or TAC(λ). *Prev. Med.* **106** 233–235. <https://doi.org/10.1016/j.ypmed.2017.10.028>
- VICIANA, J., MAYORGA-VEGA, D. and MARTÍNEZ-BAENA, A. (2016). Moderate-to-vigorous physical activity levels in physical education, school recess, and after-school time: Influence of gender, age, and weight status. *J. Phys. Act. Health* **13** 1117–1123. <https://doi.org/10.1123/jpah.2015-0537>
- WROBEL, J., ZIPUNNIKOV, V., SCHRACK, J. and GOLDSMITH, J. (2019). Registration for exponential family functional data. *Biometrics* **75** 48–57. MR3953706 <https://doi.org/10.1111/biom.12963>
- YANG, H., BALADANDAYUTHAPANI, V., RAO, A. U. K. and MORRIS, J. S. (2020). Quantile function on scalar regression analysis for distributional data. *J. Amer. Statist. Assoc.* **115** 90–106. MR4078447 <https://doi.org/10.1080/01621459.2019.1609969>
- ZHANG, J. and CHEN, H. (2022). Graph-based two-sample tests for data with repeated observations. *Statist. Sinica* **32** 391–415. MR4359638 <https://doi.org/10.5705/ss.202019.0116>
- ZHANG, J., MERIKANGAS, K. R., LI, H. and SHOU, H. (2022). Supplement to “Two-sample tests for multivariate repeated measurements of histogram objects with applications to wearable device data.” <https://doi.org/10.1214/21-AOAS1596SUPP>

PARSIMONIOUS BAYESIAN FACTOR ANALYSIS FOR MODELLING LATENT STRUCTURES IN SPECTROSCOPY DATA

BY ALESSANDRO CASA^{1,3,a}, TOM F. O’CALLAGHAN^{2,c} AND
 THOMAS BRENDAN MURPHY^{1,3,b}

¹*School of Mathematics & Statistics, University College Dublin, ^aalessandro.casa@ucd.ie, ^bbrendan.murphy@ucd.ie*

²*School of Food & Nutritional Sciences, University College Cork, ^ctom_ocalaghan@ucc.ie*

³*Vistamilk SFI Research Centre*

In recent years, within the dairy sector, animal diet and management practices have been receiving increased attention, in particular, examining the impact of pasture-based feeding strategies on the composition and quality of milk and dairy products in line with the prevalence of premium *grass-fed* dairy products appearing on market shelves. To date, methods to thoroughly investigate the more relevant differences induced by the diet on milk chemical features are limited; enhanced statistical tools exploring these differences are required.

Infrared spectroscopy techniques are widely used to collect data on milk samples and to predict milk related traits and characteristics. While these data are routinely used to predict the composition of the macro components of milk, each spectrum also provides a reservoir of unharnessed information about the sample. The accumulation and subsequent interpretation of these data present some challenges due to their high-dimensionality and the relationships amongst the spectral variables.

In this work, directly motivated by a dairy application, we propose a modification of the standard factor analysis to induce a parsimonious summary of spectroscopic data. Our proposal maps the observations into a low-dimensional latent space while simultaneously clustering the observed variables. The method indicates possible redundancies in the data, and it helps disentangle the complex relationships among the wavelengths. A flexible Bayesian estimation procedure is proposed for model fitting, providing reasonable values for the number of latent factors and clusters. The method is applied on milk mid-infrared (MIR) spectroscopy data from dairy cows on distinctly different pasture and nonpasture based diets, providing accurate modelling of the correlation, clustering of variables, and information on differences among milk samples from cows on different diets.

REFERENCES

- ABDI, H. (2007). RV coefficient and congruence coefficient. In *Encyclopedia of Measurement and Statistics* **849** 853.
- ALOTHMAN, M., HOGAN, S. A., HENNESSY, D., DILLON, P., KILCAWLEY, K. N., O’DONOVAN, M., TOBIN, J., FENELON, M. A. and O’CALLAGHAN, T. F. (2019). The “grass-fed” milk story: Understanding the impact of pasture feeding on the composition and quality of bovine milk. *Foods* **8** 350.
- ARMINGER, G. and MUTHÉN, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis–Hastings algorithm. *Psychometrika* **63** 271–300.
- BARRY, D. and HARTIGAN, J. A. (1992). Product partition models for change point problems. *Ann. Statist.* **20** 260–279. [MR1150343 https://doi.org/10.1214/aos/1176348521](https://doi.org/10.1214/aos/1176348521)
- BARTHOLOMEW, D., KNOTT, M. and MOUSTAKI, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. [MR2849614 https://doi.org/10.1002/9781119970583](https://doi.org/10.1002/9781119970583)

Key words and phrases. Dairy science, spectroscopy, chemometrics, factor analysis, redundant variables, clustering, Gibbs sampling.

- BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. [MR2806429](#) <https://doi.org/10.1093/biomet/asr013>
- BLEI, D. M. and FRAZIER, P. I. (2011). Distance dependent Chinese restaurant processes. *J. Mach. Learn. Res.* **12** 2461–2488. [MR2834504](#)
- BLUM, Y., LE MIGNON, G., LAGARRIGUE, S. and CAUSEUR, D. (2010). A factor model to analyze heterogeneity in gene expression. *BMC Bioinform.* **11** 1–12.
- BONFATTI, V., TIEZZI, F., MIGLIOR, F. and CARNIER, P. (2017). Comparison of Bayesian regression models and partial least squares regression for the development of infrared prediction equations. *J. Dairy Sci.* **100** 7306–7319. [https://doi.org/10.3168/jds.2016-12203](#)
- BOUYEYRON, C., CELEUX, G., MURPHY, T. B. and RAFTERY, A. E. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, Cambridge. [MR3967046](#) <https://doi.org/10.1017/9781108644181>
- CAPUANO, E., VAN DER VEER, G., BOERRIGTER-EENLING, R., ELGERSMA, A., RADEMAKER, J., STERIAN, A. and VAN RUTH, S. M. (2014). Verification of fresh grass feeding, pasture grazing and organic farming by cows farm milk fatty acid profile. *Food Chem.* **164** 234–241.
- CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. [MR2655722](#) <https://doi.org/10.1198/016214508000000869>
- CASA, A., O'CALLAGHAN, T. F. and MURPHY, T. B. (2022). Supplement to “Parsimonious Bayesian factor analysis for modelling latent structures in spectroscopy data.” [https://doi.org/10.1214/21-AOAS1597SUPP](#)
- DAHL, D. B., DAY, R. and TSAI, J. W. (2017). Random partition distribution indexed by pairwise information. *J. Amer. Statist. Assoc.* **112** 721–732. [MR3671765](#) <https://doi.org/10.1080/01621459.2016.1165103>
- DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2019). Multi-study factor analysis. *Biometrics* **75** 337–346. [MR3953734](#) <https://doi.org/10.1111/biom.12974>
- DE MARCHI, M., TOFFANIN, V., CASSANDRO, M. and PENASA, M. (2014). Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. *J. Dairy Sci.* **97** 1171–1186.
- DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2021). Bayesian multistudy factor analysis for high-throughput biological data. *Ann. Appl. Stat.* **15** 1723–1741. [MR4355073](#) <https://doi.org/10.1214/21-aos1456>
- DOWNEY, G. (1996). Authentication of food and food ingredients by near infrared spectroscopy. *J. Near Infrared Spectrosc.* **4** 47–61.
- DURANTE, D. (2017). A note on the multiplicative gamma process. *Statist. Probab. Lett.* **122** 198–204. [MR3584158](#) <https://doi.org/10.1016/j.spl.2016.11.014>
- ELGERSMA, A. (2012). New developments in The Netherlands: Dairies reward grazing because of public perception. *Grassl. Sci. Eur.* **17** 420–422.
- EVERITT, B. S. (1984). *An Introduction to Latent Variable Models. Monographs on Statistics and Applied Probability*. CRC Press, London. [MR0769300](#) <https://doi.org/10.1007/978-94-009-5564-6>
- FAULKNER, H., O'CALLAGHAN, T. F., MCAULIFFE, S., HENNESSY, D., STANTON, C., O'SULLIVAN, M. G., KERRY, J. P. and KILCAWLEY, K. N. (2018). Effect of different forage types on the volatile and sensory properties of bovine milk. *J. Dairy Sci.* **101** 1034–1047. [https://doi.org/10.3168/jds.2017-13141](#)
- FERRARI, F. and DUNSON, D. B. (2021). Bayesian factor analysis for inference on interactions. *J. Amer. Statist. Assoc.* **116** 1521–1532. [MR4309290](#) <https://doi.org/10.1080/01621459.2020.1745813>
- FOKOUÉ, E. and TITTERINGTON, D. (2003). Mixtures of factor analysers. Bayesian estimation and inference by stochastic simulation. *Mach. Learn.* **50** 73–94.
- 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>
- FRIZZARIN, M., O'CALLAGHAN, T. F., MURPHY, T. B., HENNESSY, D. and CASA, A. (2021). Application of machine-learning methods to milk mid-infrared spectra for discrimination of cow milk from pasture or total mixed ration diets. *J. Dairy Sci.* **104** 12394–12402. [https://doi.org/10.3168/jds.2021-20812](#)
- FRÜHWIRTH-SCHNATTER, S. (2011). Dealing with label switching under model uncertainty. In *Mixtures: Estimation and Applications. Wiley Ser. Probab. Stat.* 213–239. Wiley, Chichester. [MR2883354](#) <https://doi.org/10.1002/9781119995678.ch10>
- FRÜHWIRTH-SCHNATTER, S. and LOPES, H. F. (2010). Parsimonious Bayesian factor analysis when the number of factors is unknown. Technical report, Univ. Chicago Booth School of Business.
- GAO, C., BROWN, C. D. and ENGELHARDT, B. E. (2013). A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects. Preprint. Available at [arXiv:1310.4792](#).
- GARVEY, E. C., SANDER, T., O'CALLAGHAN, T. F., DRAKE, M., FOX, S., O'SULLIVAN, M. G., KERRY, J. P. and KILCAWLEY, K. N. (2020). A cross-cultural evaluation of liking and perception of salted butter produced from different feed systems. *Foods* **9** 1767.

- GHAHRAMANI, Z. and HINTON, G. E. (1996). The EM algorithm for mixtures of factor analyzers. Technical report CRG-TR-96-1, Univ. Toronto.
- HARTIGAN, J. A. (1990). Partition models. *Comm. Statist. Theory Methods* **19** 2745–2756. [MR1088047](#) <https://doi.org/10.1080/03610929008830345>
- HEWAVITHARANA, A. K. and VAN BRAKEL, B. (1997). Fourier transform infrared spectrometric method for the rapid determination of casein in raw milk. *Analyst* **122** 701–704.
- HIROSE, K. and KONISHI, S. (2012). Variable selection via the weighted group lasso for factor analysis models. *Canad. J. Statist.* **40** 345–361. [MR2927750](#) <https://doi.org/10.1002/cjs.11129>
- HIROSE, K. and YAMAMOTO, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Stat. Comput.* **25** 863–875. [MR3375622](#) <https://doi.org/10.1007/s11222-014-9458-0>
- HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.
- JENNRICH, R. I. and ROBINSON, S. M. (1969). A Newton–Raphson algorithm for maximum likelihood factor analysis. *Psychometrika* **34** 111–123. [MR0239703](#) <https://doi.org/10.1007/BF02290176>
- JÖRESKOG, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32** 443–482. [MR0221659](#) <https://doi.org/10.1007/BF02289658>
- KAMAL, M. and KAROUI, R. (2015). Analytical methods coupled with chemometric tools for determining the authenticity and detecting the adulteration of dairy products: A review. *Trends Food Sci. Technol.* **46** 27–48.
- KLAMI, A., VIRTANEN, S., LEPPÄÄHO, E. and KASKI, S. (2015). Group factor analysis. *IEEE Trans. Neural Netw. Learn. Syst.* **26** 2136–2147. [MR3453146](#) <https://doi.org/10.1109/TNNLS.2014.2376974>
- KNOWLES, D. and GHAHRAMANI, Z. (2007). Infinite sparse factor analysis and infinite independent components analysis. In *International Conference on Independent Component Analysis and Signal Separation* 381–388. Springer, Berlin.
- 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>
- LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14** 41–67. [MR2036762](#)
- MCPARLAND, S. and BERRY, D. P. (2016). The potential of Fourier transform infrared spectroscopy of milk samples to predict energy intake and efficiency in dairy cows. *J. Dairy Sci.* **99** 4056–4070. <https://doi.org/10.3168/jds.2015-10051>
- MCPARLAND, S., LEWIS, E., KENNEDY, E., MOORE, S. G., MCCARTHY, B., O'DONOVAN, M., BUTLER, S. T., PRYCE, J. and BERRY, D. P. (2014). Mid-infrared spectrometry of milk as a predictor of energy intake and efficiency in lactating dairy cows. *J. Dairy Sci.* **97** 5863–5871.
- MURPHY, T. B., DEAN, N. and RAFTERY, A. E. (2010). Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *Ann. Appl. Stat.* **4** 396–421. [MR2758177](#) <https://doi.org/10.1214/09-AOAS279>
- MURPHY, K., VIROLI, C. and GORMLEY, I. C. (2020). Infinite mixtures of infinite factor analysers. *Bayesian Anal.* **15** 937–963. [MR4132655](#) <https://doi.org/10.1214/19-BA1179>
- NOBILE, A. and FEARNSIDE, A. T. (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Stat. Comput.* **17** 147–162. [MR2380643](#) <https://doi.org/10.1007/s11222-006-9014-7>
- O'CALLAGHAN, T. F., HENNESSY, D., MCAULIFFE, S., KILCAWLEY, K. N., O'DONOVAN, M., DILLON, P., ROSS, R. P. and STANTON, C. (2016a). Effect of pasture versus indoor feeding systems on raw milk composition and quality over an entire lactation. *J. Dairy Sci.* **99** 9424–9440. <https://doi.org/10.3168/jds.2016-10985>
- O'CALLAGHAN, T. F., FAULKNER, H., MCAULIFFE, S., O'SULLIVAN, M. G., HENNESSY, D., DILLON, P., KILCAWLEY, K. N., STANTON, C. and ROSS, R. P. (2016b). Quality characteristics, chemical composition, and sensory properties of butter from cows on pasture versus indoor feeding systems. *J. Dairy Sci.* **99** 9441–9460. <https://doi.org/10.3168/jds.2016-11271>
- O'CALLAGHAN, T. F., MANNION, D. T., HENNESSY, D., MCAULIFFE, S., O'SULLIVAN, M. G., LEEUWENDAAL, N., BERESFORD, T. P., DILLON, P., KILCAWLEY, K. N. et al. (2017). Effect of pasture versus indoor feeding systems on quality characteristics, nutritional composition, and sensory and volatile properties of full-fat Cheddar cheese. *J. Dairy Sci.* **100** 6053–6073. <https://doi.org/10.3168/jds.2016-12508>
- O'CALLAGHAN, T. F., VÁZQUEZ-FRESNO, R., SERRA-CAYUELA, A., DONG, E., MANDAL, R., HENNESSY, D., MCAULIFFE, S., DILLON, P., WISHART, D. S. et al. (2018). Pasture feeding changes the bovine rumen and milk metabolome. *Metabolites* **8** 27.
- PAGE, G. L. and QUINTANA, F. A. (2016). Spatial product partition models. *Bayesian Anal.* **11** 265–298. [MR3465813](#) <https://doi.org/10.1214/15-BA971>
- PALLA, K., GHAHRAMANI, Z. and KNOWLES, D. (2012). A nonparametric variable clustering model. *Adv. Neural Inf. Process. Syst.* **25** 2987–2995.
- PICQUE, D., LEFIER, D., GRAPPIN, R. and CORRIEU, G. (1993). Monitoring of fermentation by infrared spectrometry: Alcoholic and lactic fermentations. *Anal. Chim. Acta* **279** 67–72.

- POURNARA, I. and WERNISCH, L. (2007). Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinform.* **8** 1–20.
- PRESS, S. J. and SHIGEMASU, K. (1989). Bayesian inference in factor analysis. In *Contributions to Probability and Statistics* 271–287. Springer, New York. [MR1024336](#)
- QUINTANA, F. A. and IGLESIAS, P. L. (2003). Bayesian clustering and product partition models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 557–574. [MR1983764](#) <https://doi.org/10.1111/1467-9868.00402>
- RAFTERY, A. E., NEWTON, M. A., SATAGOPAN, J. M. and KRIVITSKY, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In *Bayesian Statistics 8. Oxford Sci. Publ.* 371–416. Oxford Univ. Press, Oxford. [MR2433201](#)
- REID, L. M., O'DONNELL, C. P. and DOWNEY, G. (2006). Recent technological advances for the determination of food authenticity. *Trends Food Sci. Technol.* **17** 344–353.
- ROČKOVÁ, V. and GEORGE, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *J. Amer. Statist. Assoc.* **111** 1608–1622. [MR3601721](#) <https://doi.org/10.1080/01621459.2015.1100620>
- RUBIN, D. B. and THAYER, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47** 69–76. [MR0668505](#) <https://doi.org/10.1007/BF02293851>
- RUNCIE, D. E. and MUKHERJEE, S. (2013). Dissecting high-dimensional phenotypes with Bayesian sparse factor analysis of genetic covariance matrices. *Genetics* **194** 753–767.
- SCHIAVON, L. and CANALE, A. (2020). On the truncation criteria in infinite factor models. *Stat* **9** e298, 7 pp. [MR4156478](#) <https://doi.org/10.1007/s40065-018-0218-4>
- SCRUCCA, L., FOP, M., MURPHY, T. B. and RAFTERY, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8** 289.
- SONG, X.-Y. and LEE, S.-Y. (2001). Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations. *Br. J. Math. Stat. Psychol.* **54** 237–263.
- R CORE TEAM (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- WEHRHAHN, C., LEONARD, S., RODRIGUEZ, A. and XIFARA, T. (2020). A Bayesian approach to disease clustering using restricted Chinese restaurant processes. *Electron. J. Stat.* **14** 1449–1478. [MR4082474](#) <https://doi.org/10.1214/20-EJS1696>
- ZHAO, S., GAO, C., MUKHERJEE, S. and ENGELHARDT, B. E. (2016). Bayesian group factor analysis with structured sparsity. *J. Mach. Learn. Res.* **17** Paper No. 196, 47 pp. [MR3580349](#)

PHYLOGENETICALLY INFORMED BAYESIAN TRUNCATED COPULA GRAPHICAL MODELS FOR MICROBIAL ASSOCIATION NETWORKS

BY HEE CHEOL CHUNG^a, IRINA GAYNANOVA^b AND YANG NI^c

Department of Statistics, Texas A&M University, ^ahcchung@stat.tamu.edu, ^birinag@stat.tamu.edu, ^cyni@stat.tamu.edu

Microorganisms play critical roles in host health. The advancement of high-throughput sequencing technology provides opportunities for a deeper understanding of microbial interactions. However, due to the technological limitations of 16S ribosomal RNA sequencing, microbiome data are zero-inflated, and a quantitative comparison of microbial abundances cannot be made across subjects. By leveraging a recent microbiome profiling technique that quantifies 16S ribosomal RNA microbial counts, we propose a novel Bayesian graphical model that incorporates microorganisms' evolutionary history through a phylogenetic tree prior and explicitly accounts for zero inflation using the truncated Gaussian copula. Our simulation study reveals that the evolutionary information substantially improves the network estimation accuracy. We apply the proposed model to the quantitative gut microbiome data of 106 healthy subjects and identify three distinct microbial communities that are not found by existing microbial network estimation models. We further find that these communities are discriminated based on microorganisms' ability to utilize oxygen as an energy source.

REFERENCES

- ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. [MR1224394](#)
- ANBALAGAN, R., SRIKANTH, P., MANI, M., BARANI, R., SESHADRI, K. G. and JANARTHANAN, R. (2017). Next generation sequencing of oral microbiota in Type 2 diabetes mellitus prior to and after neem stick usage and correlation with serum monocyte chemoattractant-1. *Diabetes Res. Clin. Pract.* **130** 204–210. <https://doi.org/10.1016/j.diabres.2017.06.009>
- BANERJEE, O., EL GHAOU, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. [MR2417243](#)
- CANI, P. D., VAN HUL, M., LEFORT, C., DEPOMMIER, C., RASTELLI, M. and EVERARD, A. (2019). Microbial regulation of organismal energy homeostasis. *Nature Metab.* **1** 34–46.
- CHEN, Y. R., ZHENG, H. M., ZHANG, G. X., CHEN, F. L., CHEN, L. D. and YANG, Z. C. (2020a). High *Oscillospira* abundance indicates constipation and low BMI in the Guangdong Gut Microbiome Project. *Sci. Rep.* **10** 1–8.
- CHEN, L., COLLIJ, V., JAEGER, M., VAN DEN MUNCKHOF, I. C., VILA, A. V., KURILSHIKOV, A., GACESA, R., SINHA, T., OOSTING, M. et al. (2020b). Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nat. Commun.* **11** 1–12.
- CHO, I. and BLASER, M. J. (2012). The human microbiome: At the interface of health and disease. *Nat. Rev. Genet.* **13** 260–270.
- CHUNG, H. C., GAYNANOVA, I. and NI, Y. (2022). Supplement to “Phylogenetically informed Bayesian truncated copula graphical models for microbial association networks.” <https://doi.org/10.1214/21-AOAS1598SUPPA>, <https://doi.org/10.1214/21-AOAS1598SUPPB>
- CLEWELL, D. B. (1981). Plasmids, drug resistance, and gene transfer in the genus *Streptococcus*. *Microbiol. Rev.* **45** 409–436. <https://doi.org/10.1128/mr.45.3.409-436.1981>
- COOKE, F. J. and SLACK, M. P. E. (2017). 183—Gram-negative coccobacilli. In *Infectious Diseases*, 4th ed. (J. Cohen, W. G. Powderly and S. M. Opal, eds.) 1611–1627.e1. Elsevier. <https://doi.org/10.1016/B978-0-7020-6285-8.00183-0>
- DAHL, J., VANDENBERGHE, L. and ROYCHOWDHURY, V. (2008). Covariance selection for nonchordal graphs via chordal embedding. *Optim. Methods Softw.* **23** 501–520. [MR2440363](#) <https://doi.org/10.1080/10556780802102693>

- DOBRA, A. and LENKOSKI, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.* **5** 969–993. [MR2840183](#) <https://doi.org/10.1214/10-AOAS397>
- DOBRA, A., LENKOSKI, A. and RODRIGUEZ, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *J. Amer. Statist. Assoc.* **106** 1418–1433. [MR2896846](#) <https://doi.org/10.1198/jasa.2011.tm10465>
- EGLAND, P. G., PALMER, R. J. and KOLENBRANDER, P. E. (2004). Interspecies communication in *Streptococcus gordonii*–*Veillonella atypica* biofilms: Signaling in flow conditions requires juxtaposition. *Proc. Natl. Acad. Sci. USA* **101** 16917–16922.
- FAN, J., LIU, H., NING, Y. and ZOU, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 405–421. [MR3611752](#) <https://doi.org/10.1111/rssb.12168>
- FISHER, K. and PHILLIPS, C. (2009). The ecology, epidemiology and virulence of *Enterococcus*. *Microbiology* **155** 1749–1757. <https://doi.org/10.1099/mic.0.026385-0>
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.
- GARCIA-MANTRANA, I., SELMA-ROYO, M., ALCANTARA, C. and COLLADO, M. C. (2018). Shifts on gut microbiota associated to Mediterranean diet adherence and specific dietary intakes on general adult population. *Front. Microbiol.* **9** 890. <https://doi.org/10.3389/fmicb.2018.00890>
- GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
- GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. [MR3235677](#)
- GLOOR, G. B., MACKLAIM, J. M., PAWLOWSKY-GLAHN, V. and EGOZCUE, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8** 2224.
- HOFF, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.* **1** 265–283. [MR2393851](#) <https://doi.org/10.1214/07-AOAS107>
- HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. [MR1951262](#) <https://doi.org/10.1198/016214502388618906>
- JOHNSON, E. L., HEAVER, S. L., WALTERS, W. A. and LEY, R. E. (2017). Microbiome and metabolic disease: Revisiting the bacterial phylum Bacteroidetes. *J. Mol. Med.* **95** 1–8. <https://doi.org/10.1007/s00109-016-1492-2>
- KIM, M., QIE, Y., PARK, J. and KIM, C. H. (2016). Gut microbial metabolites fuel host antibody responses. *Cell Host Microbe* **20** 202–214.
- KLAASSEN, C. A. J. and WELLNER, J. A. (1997). Efficient estimation in the bivariate normal copula model: Normal margins are least favourable. *Bernoulli* **3** 55–77. [MR1466545](#) <https://doi.org/10.2307/3318652>
- KURTZ, Z. D., MÜLLER, C. L., MIRALDI, E. R., LITTMAN, D. R., BLASER, M. J. and BONNEAU, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11** e1004226. <https://doi.org/10.1371/journal.pcbi.1004226>
- KURTZ, Z. D., MÜLLER, C. L., MIRALDI, E. R., LITTMAN, D. R., BLASER, M. J. and BONNEAU, R. A. (2021). SpiecEasi: Sparse inverse covariance for ecological statistical inference. R package version 1.1.1.
- LENKOSKI, A. and DOBRA, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *J. Comput. Graph. Statist.* **20** 140–157. Supplementary material available online. [MR2816542](#) <https://doi.org/10.1198/jcgs.2010.08181>
- LEY, R. E. (2016). Prevotella in the gut: Choose carefully. *Nature Reviews Gastroenterology & Hepatology* **13** 69–70.
- LIU, H., ROEDER, K. and WASSERMAN, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. *Adv. Neural Inf. Process. Syst.* **24** 1432–1440.
- LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. [MR3059084](#) <https://doi.org/10.1214/12-AOS1037>
- LIU, F., LI, P., CHEN, M., LUO, Y., PRABHAKAR, M., ZHENG, H., HE, Y., QI, Q., LONG, H. et al. (2017). Fructooligosaccharide (FOS) and galactooligosaccharide (GOS) increase *Bifidobacterium* but reduce butyrate producing bacteria with adverse glycemic metabolism in healthy young population. *Sci. Rep.* **7** 1–12.
- LOZUPONE, C. A., STOMBAUGH, J. I., GORDON, J. I., JANSSON, J. K. and KNIGHT, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature* **489** 220–230.
- LYNCH, S. V. and PEDERSEN, O. (2016). The human intestinal microbiome in health and disease. *N. Engl. J. Med.* **375** 2369–2379.
- MA, J. (2020). Joint microbial and metabolomic network estimation with the censored Gaussian graphical model. *Stat. Biosci.* 1–22.
- MARTINEZ, K. B., PIERRE, J. F. and CHANG, E. B. (2016). The gut microbiota: The gateway to improved metabolism. *Gastroenterology Clinics* **45** 601–614.

- MARTINY, J. B., JONES, S. E., LENNON, J. T. and MARTINY, A. C. (2015). Microbiomes in light of traits: A phylogenetic perspective. *Science* **350** 9323.
- MATTHEWS, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **405** 442–451.
- MCDAVID, A., GOTTARDO, R., SIMON, N. and DRTON, M. (2019). Graphical models for zero-inflated single cell gene expression. *Ann. Appl. Stat.* **13** 848–873. [MR3963555](#) <https://doi.org/10.1214/18-AOAS1213>
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. [MR2278363](#) <https://doi.org/10.1214/009053606000000281>
- MITRA, R., MÜLLER, P., LIANG, S., YUE, L. and JI, Y. (2013). A Bayesian graphical model for ChIP-Seq data on histone modifications. *J. Amer. Statist. Assoc.* **108** 69–80. [MR3174603](#) <https://doi.org/10.1080/01621459.2012.746058>
- MULGRAVE, J. J. and GHOSAL, S. (2020). Bayesian inference in nonparanormal graphical models. *Bayesian Anal.* **15** 449–475. [MR4078721](#) <https://doi.org/10.1214/19-BA1159>
- MÜLLER, M., HERMES, G. D. A., CANFORA, E. E., SMIDT, H., MASCLÉE, A. A. M., ZOETENDAL, E. G. and BLAAK, E. E. (2020). Distal colonic transit is linked to gut microbiota diversity and microbial fermentation in humans with slow colonic transit. *Am. J. Physiol., Gastrointest. Liver Physiol.* **318** G361–G369. <https://doi.org/10.1152/ajpgi.00283.2019>
- NADERPOOR, N., MOUSA, A., GOMEZ-ARANGO, L. F., BARRETT, H. L., DEKKER NITERT, M. and DE COURTEN, B. (2019). Faecal microbiota are related to insulin sensitivity and secretion in overweight or obese adults. *Journal of Clinical Medicine* **8** 452.
- NEWMAN, M. E. and GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* **69** 026113.
- OH, J. K., CHAE, J. P., PAJARILLO, E. A. B., KIM, S. H., KWAK, M.-J., EUN, J.-S., CHEE, S. W., WHANG, K.-Y., KIM, S.-H. et al. (2020). Association between the body weight of growing pigs and the functional capacity of their gut microbiota. *J. Anim. Sci.* **91** e13418.
- OSBORNE, N., PETERSON, C. B. and VANNUCCI, M. (2022). Latent network estimation and variable selection for compositional data via variational EM. *J. Comput. Graph. Statist.* **31** 163–175. [MR4387218](#) <https://doi.org/10.1080/10618600.2021.1935971>
- PARADIS, E. (2012). *Analysis of Phylogenetics and Evolution with R*, 2nd ed. *Use R!* Springer, New York. [MR2883250](#) <https://doi.org/10.1007/978-1-4614-1743-9>
- PARADIS, E. and SCHLIEP, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. R package version 5.4.1.
- PEDROGO, D. A. M., JENSEN, M. D., VAN DYKE, C. T., MURRAY, J. A., WOODS, J. A., CHEN, J., KASHYAP, P. C. and NEHRA, V. (2018). Gut microbial carbohydrate metabolism hinders weight loss in overweight adults undergoing lifestyle intervention with a volumetric diet. In *Mayo Clinic Proceedings* **93** 1104–1110. Elsevier.
- PERALTA, G. (2016). Merging evolutionary history into species interaction networks. *Funct. Ecol.* **30** 1917–1925.
- PETERSON, C., STINGO, F. C. and VANNUCCI, M. (2015). Bayesian inference of multiple Gaussian graphical models. *J. Amer. Statist. Assoc.* **110** 159–174. [MR3338494](#) <https://doi.org/10.1080/01621459.2014.896806>
- PLUGHOEFF, K. J. and VERSALOVIC, J. (2012). Human microbiome in health and disease. *Annu. Rev. Pathol. Mech. Dis.* **7** 99–122.
- RAMAYO-CALDAS, Y., MACH, N., LEPAGE, P., LEVENEZ, F., DENIS, C., LEMONNIER, G., LEPLAT, J.-J., BILLON, Y., BERRI, M. et al. (2016). Phylogenetic network analysis applied to pig gut microbiota identifies an ecosystem structure linked with growth traits. *ISME J.* **10** 2973–2977.
- ROHR, R. P. and BASCOMPTE, J. (2014). Components of phylogenetic signal in antagonistic and mutualistic networks. *Amer. Nat.* **184** 556–564.
- ROVERATO, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Stat.* **29** 391–411. [MR1925566](#) <https://doi.org/10.1111/1467-9469.00297>
- THOMAZ, F. S., ALTEMANI, F., PANCHAL, S. K., WORRALL, S. and NITERT, M. D. (2021). The influence of wasabi on the gut microbiota of high-carbohydrate, high-fat diet-induced hypertensive Wistar rats. *J. Hum. Hypertens.* **35** 170–180.
- VANDEPUTTE, D., FALONY, G., VIEIRA-SILVA, S., TITO, R. Y., JOOSSENS, M. and RAES, J. (2016). Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* **65** 57–62. <https://doi.org/10.1136/gutjnl-2015-309618>
- VANDEPUTTE, D., KATHAGEN, G., D’HOE, K., VIEIRA-SILVA, S., VALLES-COLOMER, M., SABINO, J., WANG, J., TITO, R. Y., DE COMMER, L. et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551** 507–511.

- VAN DEN BOGERT, B., ERKUS, O., BOEKHORST, J., GOFFAU, D. M., SMID, E. J., ZOETENDAL, E. G. and KLEEREBEZEM, M. (2013). Diversity of human small intestinal *Streptococcus* and *Veillonella* populations. *FEMS Microbiol. Ecol.* **85** 376–388.
- VAN DEN BOGERT, B., MEIJERINK, M., ZOETENDAL, E. G., WELLS, J. M. and KLEEREBEZEM, M. (2014). Immunomodulatory properties of *Streptococcus* and *Veillonella* isolates from the human small intestine microbiota. *PLoS ONE* **9** e114277.
- WANG, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Anal.* **7** 867–886. [MR3000017 https://doi.org/10.1214/12-BA729](https://doi.org/10.1214/12-BA729)
- WANG, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Anal.* **10** 351–377. [MR3420886 https://doi.org/10.1214/14-BA916](https://doi.org/10.1214/14-BA916)
- WANG, S., XIAO, Y., TIAN, F., ZHAO, J., ZHANG, H., ZHAI, Q. and CHEN, W. (2020). Rational use of prebiotics for gut microbiota alterations: Specific bacterial phylotypes and related mechanisms. *J. Funct. Foods* **66** 103838.
- WASSERMAN, S., FAUST, K. et al. (1994). Social network analysis: Methods and applications.
- XIAO, J., CHEN, L., JOHNSON, S., YU, Y., ZHANG, X. and CHEN, J. (2018). Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. *Front. Microbiol.* **9** 1391.
- YANG, X., YIN, F., YANG, Y., LEPP, D., YU, H., RUAN, Z., YANG, C., YIN, Y., HOU, Y. et al. (2018). Dietary butyrate glycerides modulate intestinal microbiota composition and serum metabolites in broilers. *Sci. Rep.* **8** 1–12.
- YOON, G., CARROLL, R. J. and GAYNAKOVA, I. (2020). Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika* **107** 609–625. [MR4138979 https://doi.org/10.1093/biomet/asaa007](https://doi.org/10.1093/biomet/asaa007)
- YOON, G., GAYNAKOVA, I. and MÜLLER, C. L. (2019a). Microbial networks in SPRING—semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Front. Genet.* **10** 516. <https://doi.org/10.3389/fgene.2019.00516>
- YOON, G., GAYNAKOVA, I. and MÜLLER, C. L. (2019b). Semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. R package version 1.0.4.
- YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. [MR2367824 https://doi.org/10.1093/biomet/asm018](https://doi.org/10.1093/biomet/asm018)
- ZHANG, S. and CHEN, D.-C. (2019). Facing a new challenge: The adverse effects of antibiotics on gut microbiota and host immunity. *Chin. Med. J.* **132** 1135.
- ZHENG, J., WITTOUCK, S., SALVETTI, E., FRANZ, C. M., HARRIS, H. M., MATTARELLI, P., O'TOOLE, P. W., POT, B., VANDAMME, P. et al. (2020). A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*. *Int. J. Syst. Evol. Microbiol.* **70** 2782–2858.
- ZHOU, J., VILES, W. D., LU, B., LI, Z., MADAN, J. C., KARAGAS, M. R., GUI, J. and HOEN, A. G. (2020). Identification of microbial interaction network: Zero-inflated latent Ising model based approach. *BioData Mining* **13** 1–15.
- ZHOU, F., HE, K., LI, Q., CHAPKIN, R. S. and NI, Y. (2021). Bayesian biclustering for microbial metagenomic sequencing data via multinomial matrix factorization. *Biostatistics* **00** 1–19. <https://doi.org/10.1093/biostatistics/kxab002>
- ZOETENDAL, E. G., RAES, J., VAN DEN BOGERT, B., ARUMUGAM, M., BOOIJINK, C. C., TROOST, F. J., BORK, P., WELLS, M., DE VOS, W. M. et al. (2012). The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. *ISME J.* **6** 1415–1426.

CAUSAL INFERENCE FOR THE EFFECT OF MOBILITY ON COVID-19 DEATHS

BY MATTEO BONVINI^a, EDWARD H. KENNEDY^b, VALERIE VENTURA^c AND LARRY WASSERMAN^d

Department of Statistics and Data Science, Carnegie Mellon University, ^ambonvini@stat.cmu.edu, ^bedward@stat.cmu.edu, ^cvventura@stat.cmu.edu, ^dlarry@stat.cmu.edu

In this paper we develop statistical methods for causal inference in epidemics. Our focus is in estimating the effect of social mobility on deaths in the first year of the Covid-19 pandemic. We propose a marginal structural model motivated by a basic epidemic model. We estimate the counterfactual time series of deaths under interventions on mobility. We conduct several types of sensitivity analyses. We find that the data support the idea that reduced mobility causes reduced deaths, but the conclusion comes with caveats. There is evidence of sensitivity to model misspecification and unmeasured confounding which implies that the size of the causal effect needs to be interpreted with caution. While there is little doubt the effect is real, our work highlights the challenges in drawing causal inferences from pandemic data.

REFERENCES

- BHATT, S., FERGUSON, N., FLAXMAN, S., GANDY, A., MISHRA, S. and SCOTT, J. A. (2020). Semi-mechanistic Bayesian modeling of COVID-19 with renewal processes. arXiv preprint [arXiv:2012.00394](https://arxiv.org/abs/2012.00394).
- BJØRNSTAD, O. N. (2018). *Epidemics—Models and Data Using R. Use R!* Springer, Cham. MR3838463 <https://doi.org/10.1007/978-3-319-97487-3>
- BONVINI, M., KENNEDY, E., VENTURA, V. and WASSERMAN, L. (2021). Propensity scores and sensitivity analysis for marginal structural models with continuous treatments. In preparation.
- BONVINI, M., KENNEDY, E., VENTURA, V. and WASSERMAN, L. (2022). Supplement to “Causal inference for the effect of mobility on COVID-19 deaths.” <https://doi.org/10.1214/22-AOAS1599SUPPA>, <https://doi.org/10.1214/22-AOAS1599SUPPB>, <https://doi.org/10.1214/22-AOAS1599SUPPC>
- BRAUER, F. and CASTILLO-CHAVEZ, C. (2012). *Mathematical Models in Population Biology and Epidemiology*, 2nd ed. *Texts in Applied Mathematics* **40**. Springer, New York. MR3024808 <https://doi.org/10.1007/978-1-4614-1686-9>
- CHANG, S., PIERSON, E., KOH, P. W., GERARDIN, J., REDBIRD, B., GRUSKY, D. and LESKOVEC, J. (2020). Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 1–6.
- CHERNOZHUKOV, V., KASAHARA, H. and SCHRIMPF, P. (2021). Causal impact of masks, policies, behavior on early Covid-19 pandemic in the U.S. *J. Econometrics* **220** 23–62. MR4185124 <https://doi.org/10.1016/j.jeconom.2020.09.003>
- FONG, C., HAZLETT, C. and IMAI, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *Ann. Appl. Stat.* **12** 156–177. MR3773389 <https://doi.org/10.1214/17-AOAS1101>
- IHME (2020). Modeling COVID-19 scenarios for the United States. *Nat. Med.*
- KERMACK, W. O. and MCKENDRICK, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A, Math. Phys. Sci.* **115** 700–721.
- LIAO, X. and MEYER, M. (2018). cgam: Constrained generalized additive model.
- MEYER, M. C. (2008). Inference using shape-restricted regression splines. *Ann. Appl. Stat.* **2** 1013–1033. MR2516802 <https://doi.org/10.1214/08-AOAS167>
- MEYER, M. C. (2018). A framework for estimation and inference in generalized additive models with shape and order restrictions. *Statist. Sci.* **33** 595–614. MR3881210 <https://doi.org/10.1214/18-STS671>
- NEUGEBAUER, R. and VAN DER LAAN, M. (2007). Nonparametric causal effects based on marginal structural models. *J. Statist. Plann. Inference* **137** 419–434. MR2298947 <https://doi.org/10.1016/j.jspi.2005.12.008>
- NEWBY, W. K. and WEST, K. D. (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55** 703–708. MR0890864 <https://doi.org/10.2307/1913610>

- ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512. [MR0877758](#) [https://doi.org/10.1016/0270-0255\(86\)90088-6](https://doi.org/10.1016/0270-0255(86)90088-6)
- ROBINS, J. M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS* 113–159.
- ROBINS, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (Minneapolis, MN, 1997). *IMA Vol. Math. Appl.* **116** 95–133. Springer, New York. [MR1731682](#) https://doi.org/10.1007/978-1-4612-1284-3_2
- ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology.
- ROBINS, J. M. and WASSERMAN, L. (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence* 409–420. Morgan Kaufmann.
- 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>
- SCOTT, J. A., GANDY, A., MISHRA, S., BHATT, S., FLAXMAN, S., UNWIN, H. J. T. and ISH-HOROWICZ, J. (2021). Epidemia: An R package for semi-mechanistic bayesian modelling of infectious diseases using point processes. arXiv preprint [arXiv:2110.12461](#).
- SHI, Y. and BAN, X. (2020). Capping mobility to control COVID-19: A collision-based infectious disease transmission model. MedRxiv.
- TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data. Springer Series in Statistics*. Springer, New York. [MR2233926](#)
- UNWIN, H. J. T., MISHRA, S., BRADLEY, V. C., GANDY, A., MELLAN, T. A., COUPLAND, H., ISH-HOROWICZ, J., VOLLMER, M. A., WHITTAKER, C. et al. (2020). State-level tracking of COVID-19 in the United States. *Nat. Commun.* **11** 1–9.
- VANSTEELENDT, S. and JOFFE, M. (2014). Structural nested models and G-estimation: The partially realized promise. *Statist. Sci.* **29** 707–731. [MR3300367](#) <https://doi.org/10.1214/14-STS493>
- XIONG, C., HU, S., YANG, M., LUO, W. and ZHANG, L. (2020). Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proc. Natl. Acad. Sci. USA* **117** 27087–27089.
- YADLOWSKY, S., NAMKOONG, H., BASU, S., DUCHI, J. and TIAN, L. (2018). Bounds on the conditional and average treatment effect with unobserved confounding factors. arXiv preprint [arXiv:1808.09521](#).
- ZHAO, Q., SMALL, D. S. and BHATTACHARYA, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 735–761. [MR3997099](#)
- ZHOU, X. and WODTKE, G. T. (2018). Residual balancing weights for marginal structural models: With application to analyses of time-varying treatments and causal mediation. arXiv preprint [arXiv:1807.10869](#).

BAYESIAN HIERARCHICAL RANDOM-EFFECTS META-ANALYSIS AND DESIGN OF PHASE I CLINICAL TRIALS

BY RUITAO LIN^{1,a}, HAOLUN SHI^{2,d}, GUOSHENG YIN^{3,e}, PETER F. THALL^{1,b},
 YING YUAN^{1,c} AND CHRISTOPHER R. FLOWERS^{4,f}

¹Department of Biostatistics, The University of Texas MD Anderson Cancer Center, ^arlin@mdanderson.org,
^brex@mdanderson.org, ^cyyuan@mdanderson.org

²Department of Statistics and Actuarial Science, Simon Fraser University, ^dhaoluns@sfu.ca

³Department of Statistics and Actuarial Science, The University of Hong Kong, ^egyin@hku.hk

⁴Department of Lymphoma/Myeloma, The University of Texas MD Anderson Cancer Center, ^fcrflowers@mdanderson.org

We propose a curve-free random-effects meta-analysis approach to combining data from multiple phase I clinical trials to identify an optimal dose. Our method accounts for between-study heterogeneity that may stem from different study designs, patient populations, or tumor types. We also develop a meta-analytic-predictive (MAP) method, based on a power prior, that incorporates data from multiple historical studies into the design and conduct of a new phase I trial. Performances of the proposed methods for data analysis and trial design are evaluated by extensive simulation studies. The proposed random-effects meta-analysis method provides more reliable dose selection than comparators that rely on parametric assumptions. The MAP-based dose-finding designs are generally more efficient than those that do not borrow information, especially when the current and historical studies are similar. The proposed methodologies are illustrated by a meta-analysis of five historical phase I studies of Sorafenib and design of a new phase I trial.

REFERENCES

- AITCHISON, J. and SHEN, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika* **67** 261–272. [MR0581723 https://doi.org/10.2307/2335470](https://doi.org/10.2307/2335470)
- AWADA, A., HENDLISZ, A., GIL, T. et al. (2005). Phase I safety and pharmacokinetics of BAY 43-9006 administered for 21 days on/7 days off in patients with advanced, refractory solid tumours. *Br. J. Cancer* **92** 1855–1861.
- BABB, J., ROGATKO, A. and ZACKS, S. (1998). Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Stat. Med.* **17** 1103–1120.
- BORTHAKUR, G., KANTARIAN, H., RAVANDI, F. et al. (2011). Phase 1 study of sorafenib in patients with refractory or relapsed acute leukemias. *Haematol.* **96** 61–68.
- BOUCHER, M. and BENNETTS, M. (2015). The many flavors of model-based meta-analysis: Part I—introduction and landmark data. *CPT: Pharmacometrics & Systems Pharmacology* **5** 54–64.
- BRIL, G., DYKSTRA, R., PILLERS, C. and ROBERTSON, T. (1984). Isotonic regression in two independent variables. *Appl. Stat.* **33** 352–357.
- CHEN, M.-H. and IBRAHIM, J. G. (2006). The relationship between the power prior and hierarchical models. *Bayesian Anal.* **1** 551–574. [MR2221288 https://doi.org/10.1214/06-BA118](https://doi.org/10.1214/06-BA118)
- CHEN, M.-H., IBRAHIM, J. G., LAM, P., YU, A. and ZHANG, Y. (2011). Bayesian design of noninferiority trials for medical devices using historical data. *Biometrics* **67** 1163–1170. [MR2829252 https://doi.org/10.1111/j.1541-0420.2011.01561.x](https://doi.org/10.1111/j.1541-0420.2011.01561.x)
- CHOW, S.-C., CHIANG, C., LIU, J. and HSIAO, C.-F. (2012). Statistical methods for bridging studies. *J. Biopharm. Statist.* **22** 903–915. [MR2970563 https://doi.org/10.1080/10543406.2012.701578](https://doi.org/10.1080/10543406.2012.701578)
- CLARK, J. W., EDER, J. P., RYAN, D. et al. (2005). Safety and pharmacokinetics of the dual action Raf kinase and vascular endothelial growth factor receptor inhibitor, BAY 43-9006, in patients with advanced, refractory solid tumors. *Clin. Cancer Res.* **11** 5472–5480.

- CUNANAN, K. M. and KOOPMEINERS, J. S. (2018). Hierarchical models for sharing information across populations in phase I dose-escalation studies. *Stat. Methods Med. Res.* **27** 3447–3459. [MR3865285](#) <https://doi.org/10.1177/0962280217703812>
- GARCÍA, V. M., OLMOS, D., GOMEZ-ROCA, C. et al. (2014). Dose-response relationship in phase I clinical trials: A European Drug Development Network (EDDN) collaboration study. *Clin. Cancer Res.* **20** 5663–5671.
- GASPARINI, M. and EISELE, J. (2000). A curve-free method for phase I clinical trials. *Biometrics* **56** 609–615. [MR1795024](#) <https://doi.org/10.1111/j.0006-341X.2000.00609.x>
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. [MR2221284](#) <https://doi.org/10.1214/06-BA117A>
- GEZMU, M. and FLOURNOY, N. (2006). Group up-and-down designs for dose-finding. *J. Statist. Plann. Inference* **136** 1749–1764. [MR2255594](#) <https://doi.org/10.1016/j.jspi.2005.08.002>
- GOULD, A. L., JIN, T., ZHANG, L. X. and WANG, W. W. B. (2012). A predictive Bayesian approach to the design and analysis of bridging studies. *J. Biopharm. Statist.* **22** 916–934. [MR2970564](#) <https://doi.org/10.1080/10543406.2012.701579>
- HIGGINS, J. P. T., THOMPSON, S. G. and SPIEGELHALTER, D. J. (2009). A re-evaluation of random-effects meta-analysis. *J. Roy. Statist. Soc. Ser. A* **172** 137–159. [MR2655609](#) <https://doi.org/10.1111/j.1467-985X.2008.00552.x>
- HOBBS, B. P., CARLIN, B. P. and SARGENT, D. J. (2013). Adaptive adjustment of the randomization ratio using historical control data. *Clin. Trials* **10** 430–440.
- HOBBS, B. P., SARGENT, D. J. and CARLIN, B. P. (2012). Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Anal.* **7** 639–673. [MR2981631](#) <https://doi.org/10.1214/12-BA722>
- HOBBS, B. P., CARLIN, B. P., MANDREKAR, S. J. and SARGENT, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* **67** 1047–1056. [MR2829239](#) <https://doi.org/10.1111/j.1541-0420.2011.01564.x>
- HUANG, S. M. and TEMPLE, R. (2008). Is this the drug or dose for you?: Impact and consideration of ethnic factors in global drug development, regulatory review, and clinical practice. *Clin. Pharmacol. Ther.* **84** 287–294.
- IBRAHIM, J. G. and CHEN, M.-H. (2000). Power prior distributions for regression models. *Statist. Sci.* **15** 46–60. [MR1842236](#) <https://doi.org/10.1214/ss/1009212673>
- IBRAHIM, J. G., CHEN, M.-H., XIA, H. A. and LIU, T. (2012). Bayesian meta-experimental design: Evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. *Biometrics* **68** 578–586. [MR2959625](#) <https://doi.org/10.1111/j.1541-0420.2011.01679.x>
- LEE, S. M. and CHEUNG, Y. K. (2009). Model calibration in the continual reassessment method. *Clin. Trials* **6** 227–238.
- LEE, J., THALL, P. F., JI, Y. and MÜLLER, P. (2015). Bayesian dose-finding in two treatment cycles based on the joint utility of efficacy and toxicity. *J. Amer. Statist. Assoc.* **110** 711–722. [MR3367259](#) <https://doi.org/10.1080/01621459.2014.926815>
- LENK, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.* **83** 509–516. [MR0971380](#)
- LIN, R. and YIN, G. (2017). Nonparametric overdose control with late-onset toxicity in phase I clinical trials. *Biostatistics* **18** 180–194. [MR3612282](#) <https://doi.org/10.1093/biostatistics/kxw038>
- LIN, R., SHI, H., YIN, G., THALL, P. F., YUAN, Y. and FLOWERS, C. R. (2022). Supplement to “Bayesian hierarchical random-effects meta-analysis and design of phase I clinical trials.” <https://doi.org/10.1214/22-AOAS1600SUPP>
- LIU, S. and JOHNSON, V. E. (2016). A robust Bayesian dose-finding design for phase I/II clinical trials. *Biostatistics* **17** 249–263. [MR3515998](#) <https://doi.org/10.1093/biostatistics/kxv040>
- LIU, S. and YUAN, Y. (2015). Bayesian optimal interval designs for phase I clinical trials. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 507–523. [MR3325461](#) <https://doi.org/10.1111/rssc.12089>
- LIU, S., PAN, H., XIA, J., HUANG, Q. and YUAN, Y. (2015). Bridging continual reassessment method for phase I clinical trials in different ethnic populations. *Stat. Med.* **34** 1681–1694. [MR3334684](#) <https://doi.org/10.1002/sim.6442>
- LU, G. and ADES, A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Stat. Med.* **23** 3105–3124.
- MINAMI, H., KAWADA, K., EBI, H. et al. (2008). Phase I and pharmacokinetic study of sorafenib, an oral multikinase inhibitor, in Japanese patients with advanced refractory solid tumors. *Cancer Sci.* **99** 1492–1498.
- MOORE, M., HIRTE, H. W., SIU, L. et al. (2005). Phase I study to determine the safety and pharmacokinetics of the novel Raf kinase and VEGFR inhibitor BAY 43-9006, administered for 28 days on/7 days off in patients with advanced, refractory solid tumors. *Ann. Oncol.* **16** 1688–1694.

- MORITA, S. (2011). Application of the continual reassessment method to a phase I dose-finding trial in Japanese patients: East meets West. *Stat. Med.* **30** 2090–2097. [MR2829159](#) <https://doi.org/10.1002/sim.3999>
- MORITA, S., THALL, P. F. and MÜLLER, P. (2008). Determining the effective sample size of a parametric prior. *Biometrics* **64** 595–602, 669–670. [MR2432433](#) <https://doi.org/10.1111/j.1541-0420.2007.00888.x>
- MOULD, D. R. (2012). Model-based meta-analysis: An important tool for making quantitative decisions during drug development. *Clin. Pharmacol. Ther.* **92** 283–286.
- NEUENSCHWANDER, B., BRANSON, M. and GSPONER, T. (2008). Critical aspects of the Bayesian approach to phase I cancer trials. *Stat. Med.* **27** 2420–2439. [MR2432497](#) <https://doi.org/10.1002/sim.3230>
- NEUENSCHWANDER, B., CAPKUN-NIGGLI, G., BRANSON, M. and SPIEGELHALTER, D. J. (2010). Summarizing historical information on controls in clinical trials. *Clin. Trials* **7** 5–18. <https://doi.org/10.1177/1740774509356002>
- NEUENSCHWANDER, B., WEBER, S., SCHMIDLI, H. and O'HAGAN, A. (2020). Predictively consistent prior effective sample sizes. *Biometrics* **76** 578–587. <https://doi.org/10.1111/biom.13252>
- O'QUIGLEY, J., PEPE, M. and FISHER, L. (1990). Continual reassessment method: A practical design for phase I clinical trials in cancer. *Biometrics* **46** 33–48. [MR1059105](#) <https://doi.org/10.2307/2531628>
- SCHMIDLI, H., GSTEIGER, S., ROYCHOUDHURY, S., O'HAGAN, A., SPIEGELHALTER, D. and NEUENSCHWANDER, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* **70** 1023–1032. [MR3295763](#) <https://doi.org/10.1111/biom.12242>
- SPIEGELHALTER, D. J., ABRAMS, K. R. and MYLES, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, Chichester.
- STANGL, D. and BERRY, D. A. (2000). *Meta-Analysis in Medicine and Health Policy*. CRC Press, Boca Raton.
- STORER, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics* **45** 925–937. [MR1029610](#) <https://doi.org/10.2307/2531693>
- STRUMBERG, D., RICHLI, H., HILGER, R. A. et al. (2005). Phase I clinical and pharmacokinetic study of the Novel Raf kinase and vascular endothelial growth factor receptor inhibitor BAY 43-9006 in patients with advanced refractory solid tumors. *J. Clin. Oncol.* **23** 965–972.
- URSINO, M., RÖVER, C., ZOHAR, S. and FRIEDE, T. (2021). Random-effects meta-analysis of phase I dose-finding studies using stochastic process priors. *Ann. Appl. Stat.* **15** 174–193. [MR4255273](#) <https://doi.org/10.1214/20-aos1390>
- YASUDA, S. U., ZHANG, L. and HUANG, S. M. (2008). The role of ethnicity in variability in response to drugs: Focus on clinical pharmacology studies. *Clin. Pharmacol. Ther.* **84** 417–423.
- YIN, G. and IBRAHIM, J. G. (2005). Cure rate models: A unified approach. *Canad. J. Statist.* **33** 559–570. [MR2232380](#) <https://doi.org/10.1002/cjs.5550330407>
- YIN, G., LI, Y. and JI, Y. (2006). Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios. *Biometrics* **62** 777–784. [MR2247206](#) <https://doi.org/10.1111/j.1541-0420.2006.00534.x>
- YUAN, Z. and CHAPPELL, R. (2004). Isotonic designs for phase I cancer clinical trials with multiple risk groups. *Clin. Trials* **1**, 499–508.
- YUAN, Y., HESS, K. R., HILSENBECK, S. G. and GILBERT, M. R. (2016). Bayesian optimal interval design: A simple and well-performing design for phase I oncology trials. *Clin. Cancer Res.* **22** 4291–4301.
- ZOHAR, S., KATSAHIAN, S. and O'QUIGLEY, J. (2011). An approach to meta-analysis of dose-finding studies. *Stat. Med.* **30** 2109–2116. [MR2829161](#) <https://doi.org/10.1002/sim.4121>

ESTIMATING FUNCTIONAL PARAMETERS FOR UNDERSTANDING THE IMPACT OF WEATHER AND GOVERNMENT INTERVENTIONS ON COVID-19 OUTBREAK

BY CHIH-LI SUNG^a

Department of Statistics and Probability, Michigan State University, ^asungchih@msu.edu

As the coronavirus disease 2019 (COVID-19) has shown profound effects on public health and the economy worldwide, it becomes crucial to assess the impact on the virus transmission and develop effective strategies to address the challenge. A new statistical model, derived from the SIR epidemic model with functional parameters, is proposed to understand the impact of weather and government interventions on the virus spread in the presence of asymptomatic infections among eight metropolitan areas in the United States. The model uses Bayesian inference with Gaussian process priors to study the functional parameters nonparametrically, and sensitivity analysis is adopted to investigate the main and interaction effects of these factors. This analysis reveals several important results, including the potential interaction effects between weather and government interventions, which shed new light on the effective strategies for policymakers to mitigate the COVID-19 outbreak.

REFERENCES

- ALMUKHTAR, S., AUFRICHTIG, A., BARNARD, A., BLOCH, M., CAI, W., CALDERONE, J., COLLINS, K., CONLEN, M., COOK, L. et al. (2020). Coronavirus in the US: Latest map and case count. *The New York Times*.
- AMBROSIO, B. and AZIZ-ALAOU, M. (2020). On a coupled time-dependent SIR models fitting with New York and New-Jersey states COVID-19 data. *Biology (Basel)* **9** 135.
- ANSUMALI, S., KAUSHAL, S., KUMAR, A., PRAKASH, M. K. and VIDYASAGAR, M. (2020). Modelling a pandemic with asymptomatic patients, impact of lockdown and herd immunity, with applications to SARS-CoV-2. *Annu. Rev. Control* **50** 432–447. [MR4188919 https://doi.org/10.1016/j.arcontrol.2020.10.003](https://doi.org/10.1016/j.arcontrol.2020.10.003)
- BANERJEE, S. and GELFAND, A. E. (2002). Prediction, interpolation and regression for spatially misaligned data. *Sankhyā Ser. A* **64** 227–245. [MR1981755](https://doi.org/10.1016/j.sankhyā.2002.04.003)
- BROWN, D. A. and ATAMURKTUR, S. (2018). Nonparametric functional calibration of computer models. *Statist. Sinica* **28** 721–742. [MR3791085](https://doi.org/10.1016/j.sankhyā.2018.04.003)
- CAFLISCH, R. E. (1998). Monte Carlo and quasi-Monte Carlo methods. *Acta Numer.* **7** 1–49. [MR1689431 https://doi.org/10.1017/S0962492900002804](https://doi.org/10.1017/S0962492900002804)
- CARSON, R. T., CARSON, S. L., DYE, T. K., MAYFIELD, S. L., MOYER, D. C. and CHU, A. Y. (2020). COVID-19's US temperature response profile. *MedRxiv*.
- CAUCHEMEZ, S., NOUVELLET, P., CORI, A., JOMBART, T., GARSKE, T., CLAPHAM, H., MOORE, S., MILLS, H. L., SALJE, H. et al. (2016). Unraveling the drivers of MERS-CoV transmission. *Proc. Natl. Acad. Sci. USA* **113** 9081–9086.
- CHEN, Y.-C., LU, P.-E., CHANG, C.-S. and LIU, T.-H. (2020). A time-dependent SIR model for COVID-19 with undetectable infected persons. *IEEE Trans. Netw. Sci. Eng.* **7** 3279–3294. [MR4205631](https://doi.org/10.1109/NETSC.2020.3000000)
- COOPER, I., MONDAL, A. and ANTONOPOULOS, C. G. (2020). A SIR model assumption for the spread of COVID-19 in different communities. *Chaos Solitons Fractals* **139** 110057, 14. [MR4123704 https://doi.org/10.1016/j.chaos.2020.110057](https://doi.org/10.1016/j.chaos.2020.110057)
- COWLING, B. J., HO, L. M. and LEUNG, G. M. (2008). Effectiveness of control measures during the SARS epidemic in Beijing: A comparison of the Rt curve and the epidemic curve. *Epidemiol. Infect.* **136** 562–566.
- COWLING, B. J., ALI, S. T., NG, T. W., TSANG, T. K., LI, J. C., FONG, M. W., LIAO, Q., KWAN, M. Y., LEE, S. L. et al. (2020). Impact assessment of non-pharmaceutical interventions against COVID-19 and influenza in Hong Kong: An observational study. *Lancet Public Health* **5** e279–e288.

- D'ARIENZO, M. and CONIGLIO, A. (2020). Assessment of the SARS-CoV-2 basic reproduction number, R_0 , based on the early phase of COVID-19 outbreak in Italy. *Biosafety and Health*.
- DIEKMANN, O., HEESTERBEEK, H. and BRITTON, T. (2013). *Mathematical Tools for Understanding Infectious Disease Dynamics*. Princeton Series in Theoretical and Computational Biology. Princeton Univ. Press, Princeton, NJ. **MR3015083**
- DIETZ, K. (1993). The estimation of the basic reproduction number for infectious diseases. *Stat. Methods Med. Res.* **2** 23–41. <https://doi.org/10.1177/096228029300200103>
- FLAXMAN, S., MISHRA, S., GANDY, A., UNWIN, H. J. T., MELLAN, T. A., COUPLAND, H., WHITTAKER, C., ZHU, H., BERAH, T. et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584** 257–261.
- FRICKER, T. E., OAKLEY, J. E. and URBAN, N. M. (2013). Multivariate Gaussian process emulators with nonseparable covariance structures. *Technometrics* **55** 47–56. **MR3038484** <https://doi.org/10.1080/00401706.2012.715835>
- GUPTA, A., PRADHAN, B. and MAULUD, K. N. A. (2020). Estimating the impact of daily weather on the temporal pattern of COVID-19 outbreak in India. *Earth Systems and Environment* **4** 523–534.
- HALDAR, A. and SETHI, N. (2020). The effect of country-level factors and government intervention on the incidence of COVID-19. *Asian Economics Letters* **1** 17804.
- HAUG, N., GEYRHOFFER, L., LONDEI, A., DERIC, E., DESVARS-LARRIVE, A., LORETO, V., PINIOR, B., THURNER, S. and KLIMEK, P. (2020). Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat. Hum. Behav.* 1–10.
- HE, X., LAU, E. H., WU, P., DENG, X., WANG, J., HAO, X., LAU, Y. C., WONG, J. Y., GUAN, Y. et al. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26** 672–675.
- HEESTERBEEK, H., ANDERSON, R. M., ANDREASEN, V., BANSAL, S., DE ANGELIS, D., DYE, C., EAMES, K. T., EDMUNDS, W. J., FROST, S. D. et al. (2015). Modeling infectious disease dynamics in the complex landscape of global health. *Science* **347** aaa4339.
- HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* **19** 293–325. **MR0026294** <https://doi.org/10.1214/aoms/1177730196>
- HONG, H. G. and LI, Y. (2020). Estimation of time-varying reproduction numbers underlying epidemiological processes: A new statistical tool for the COVID-19 pandemic. *PLoS ONE* **15** e0236464.
- JAMIL, T., ALAM, I., GOJOBORI, T. and DUARTE, C. M. (2020). No evidence for temperature-dependence of the COVID-19 epidemic. *Frontiers in Public Health* **8** 436.
- 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>
- KERMACK, W. O. and MCKENDRICK, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **115** 700–721.
- LE GRATIET, L., CANNAMELA, C. and IOOSS, B. (2014). A Bayesian approach for global sensitivity analysis of (multifidelity) computer codes. *SIAM/ASA J. Uncertain. Quantificat.* **2** 336–363. **MR3283912** <https://doi.org/10.1137/130926869>
- LEE, J. C., MERVOSH, S., AVILA, Y., HARVEY, B., MATTHEWS, A. L., GAMIO, L., POPOVICH, N., MCCANN, A. and KEEFE, J. (2020). See coronavirus restrictions and mask mandates for all 50 states. *The New York Times*.
- LI, R., PEI, S., CHEN, B., SONG, Y., ZHANG, T., YANG, W. and SHAMAN, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368** 489–493.
- LIU, Y., YAN, L.-M., WAN, L., XIANG, T.-X., LE, A., LIU, J.-M., PEIRIS, M., POON, L. L. and ZHANG, W. (2020). Viral dynamics in mild and severe cases of COVID-19. *Lancet Infect. Dis.* **20** 656–657.
- 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>
- NOH, J. and DANUSER, G. (2021). Estimation of the fraction of COVID-19 infected people in US states and countries worldwide. *PLoS ONE* **16** e0246772.
- OSTHUS, D., HICKMANN, K. S., CARAGEA, P. C., HIGDON, D. and DEL VALLE, S. Y. (2017). Forecasting seasonal influenza with a state-space SIR model. *Ann. Appl. Stat.* **11** 202–224. **MR3634321** <https://doi.org/10.1214/16-AOAS1000>
- PEI, S., YAMANA, T. K., KANDULA, S., GALANTI, M. and SHAMAN, J. (2021). Burden and characteristics of COVID-19 in the United States during 2020. *Nature* to appear.
- PIAZZOLA, C., TAMELLINI, L. and TEMPONE, R. (2021). A note on tools for prediction under uncertainty and identifiability of SIR-like dynamical systems for epidemiology. *Math. Biosci.* **332** Paper No. 108514, 21. **MR4205198** <https://doi.org/10.1016/j.mbs.2020.108514>

- PLUMLEE, M., JOSEPH, V. R. and YANG, H. (2016). Calibrating functional parameters in the ion channel models of cardiac cells. *J. Amer. Statist. Assoc.* **111** 500–509. [MR3538682](#) <https://doi.org/10.1080/01621459.2015.1119695>
- QIAN, P. Z. G., WU, H. and WU, C. F. J. (2008). Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics* **50** 383–396. [MR2457574](#) <https://doi.org/10.1198/004017008000000262>
- R CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR2514435](#)
- ROBINSON, M. and STILIANAKIS, N. I. (2013). A model for the emergence of drug resistance in the presence of asymptomatic infections. *Math. Biosci.* **243** 163–177. [MR3065204](#) <https://doi.org/10.1016/j.mbs.2013.03.003>
- RODA, W. C., VARUGHESE, M. B., HAN, D. and LI, M. Y. (2020). Why is it difficult to accurately predict the COVID-19 epidemic? *Infect. Dis. Model.* **5** 271–281.
- 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](#)
- SOBOL', I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput. Exp.* **1** 407–414 (1995). [MR1335161](#)
- SOUICY, J.-P. R., STURROCK, S. L., BERRY, I., DANEMAN, N., MACFADDEN, D. R. and BROWN, K. A. (2020). Estimating the effect of physical distancing on the COVID-19 pandemic using an urban mobility index. MedRxiv.
- SUNG, C. L. (2022). Supplement to “Estimating functional parameters for understanding the impact of weather and government interventions on COVID-19 outbreak.” <https://doi.org/10.1214/22-AOAS1601SUPPA>, <https://doi.org/10.1214/22-AOAS1601SUPPB>
- SUNG, C.-L. and HUNG, Y. (2020). Efficient calibration for imperfect epidemic models with applications to the analysis of COVID-19. arXiv preprint [arXiv:2009.12523](https://arxiv.org/abs/2009.12523).
- SVENSON, J. and SANTNER, T. (2016). Multiobjective optimization of expensive-to-evaluate deterministic computer simulator models. *Comput. Statist. Data Anal.* **94** 250–264. [MR3412823](#) <https://doi.org/10.1016/j.csda.2015.08.011>
- TUO, R. and WU, C. F. J. (2015). Efficient calibration for imperfect computer models. *Ann. Statist.* **43** 2331–2352. [MR3405596](#) <https://doi.org/10.1214/15-AOS1314>
- U. S. CENTERS FOR DISEASE CONTROL AND PREVENTION (2021). Estimated disease burden of COVID-19. Available at <https://www.cdc.gov/coronavirus/2019-ncov/>.
- WANG, G., ZHANG, Y., ZHAO, J., ZHANG, J. and JIANG, F. (2020). Mitigate the effects of home confinement on children during the COVID-19 outbreak. *Lancet* **395** 945–947.
- THE WEATHER COMPANY (2020). Weather underground. Available at <https://www.wunderground.com/>.
- WELCH, W. J., BUCK, R. J., SACKS, J., WYNN, H. P., MITCHELL, T. J. and MORRIS, M. D. (1992). Screening, predicting, and computer experiments. *Technometrics* **34** 15–25.
- WILSON, D. J. (2020). Weather, social distancing, and the spread of COVID-19. MedRxiv.
- WÖLFEL, R., CORMAN, V. M., GUGGEMOS, W., SEILMAIER, M., ZANGE, S., MÜLLER, M. A., NIEMEYER, D., JONES, T. C., VOLLMAR, P. et al. (2020). Virological assessment of hospitalized patients with COVID-2019. *Nature* **581** 465–469.
- XU, R., RAHMADAD, H., GUPTA, M., DIGENNARO, C., GHAFARZADEGAN, N., AMINI, H. and JALALI, M. S. (2020). The modest impact of weather and air pollution on COVID-19 transmission. MedRxiv.
- YU, X. (2020). Impact of mitigating interventions and temperature on the instantaneous reproduction number in the COVID-19 epidemic among 30 US metropolitan areas. MedRxiv.
- ZHANG, J., LITVINOVA, M., WANG, W., WANG, Y., DENG, X., CHEN, X., LI, M., ZHENG, W., YI, L., CHEN, X. et al. (2020). Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: A descriptive and modelling study. *Lancet Infect. Dis.* **20** 793–802.
- ZHAO, S., LIN, Q., RAN, J., MUSA, S. S., YANG, G., WANG, W., LOU, Y., GAO, D., YANG, L., HE, D. and WANG, M. H. (2020). Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *Int. J. Infect. Dis.* **92** 214–217.

CLUSTERING AND FORECASTING MULTIPLE FUNCTIONAL TIME SERIES

BY CHEN TANG^{1,a}, HAN LIN SHANG^{2,c} AND YANRONG YANG^{1,b}

¹Research school of Finance, Actuarial Studies and Statistics, The Australian National University, ^achen.tang@anu.edu.au,
^byanrong.yang@anu.edu.au

²Department of Actuarial Studies and Business Analytics, Macquarie University, ^chanlin.shang@mq.edu.au

Modeling and forecasting homogeneous age-specific mortality rates of multiple countries could lead to improvements in long-term forecasting. Data fed into joint models are often grouped according to nominal attributes, such as geographic regions, ethnic groups, and socioeconomic status, which may still contain heterogeneity and deteriorate the forecast results. Our paper proposes a novel clustering technique to pursue homogeneity among multiple functional time series, based on functional panel data modeling, to address this issue. Using a functional panel data model with fixed effects, we can extract common functional time series features. These common features could be decomposed into two components: the functional time trend and the mode of variations of functions (functional pattern). The functional time trend reflects the dynamics across time, while the functional pattern captures the fluctuations within curves. The proposed clustering method searches for homogeneous age-specific mortality rates of multiple countries by accounting for both the modes of variations and the temporal dynamics among curves. We demonstrate that the proposed clustering technique outperforms other existing methods through a Monte Carlo simulation and could handle complicated cases with slow decaying eigenvalues. In empirical data analysis we find that the clustering results of age-specific mortality rates can be explained by the combination of geographic region, ethnic groups, and socioeconomic status. We further show that our model produces more accurate forecasts than several benchmark methods in forecasting age-specific mortality rates.

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>
- ANDREWS, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59** 817–858. [MR1106513](#) <https://doi.org/10.2307/2938229>
- AUE, A., NORINHO, D. D. and HÖRMANN, S. (2015). On the prediction of stationary functional time series. *J. Amer. Statist. Assoc.* **110** 378–392. [MR3338510](#) <https://doi.org/10.1080/01621459.2014.909317>
- BAI, J. (2009). Panel data models with interactive fixed effects. *Econometrica* **77** 1229–1279. [MR2547073](#) <https://doi.org/10.3982/ECTA6135>
- BOIVIN, J. and NG, S. (2006). Are more data always better for factor analysis? *J. Econometrics* **132** 169–194. [MR2271395](#) <https://doi.org/10.1016/j.jeconom.2005.01.027>
- BOOTH, H. and TICKLE, L. (2008). Mortality modelling and forecasting: A review of methods. *Ann. Actuar. Sci.* **3** 3–43.
- BOUYEYRON, C., CÔME, E. and JACQUES, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *Ann. Appl. Stat.* **9** 1726–1760. [MR3456352](#) <https://doi.org/10.1214/15-AOAS861>
- BOUYEYRON, C., GIRARD, S. and SCHMID, C. (2007). High-dimensional data clustering. *Comput. Statist. Data Anal.* **52** 502–519. [MR2409998](#) <https://doi.org/10.1016/j.csda.2007.02.009>
- BOUYEYRON, C. and JACQUES, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Adv. Data Anal. Classif.* **5** 281–300. [MR2860102](#) <https://doi.org/10.1007/s11634-011-0095-6>

Key words and phrases. Functional panel data, multilevel functional data, functional time series, functional principal component analysis, age-specific mortality forecasting.

- CHIOU, J.-M. (2012). Dynamical functional prediction and classification, with application to traffic flow prediction. *Ann. Appl. Stat.* **6** 1588–1614. [MR3058676](#) <https://doi.org/10.1214/12-AOAS595>
- CHIOU, J.-M. and LI, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 679–699. [MR2370075](#) <https://doi.org/10.1111/j.1467-9868.2007.00605.x>
- CHIOU, J.-M. and MÜLLER, H.-G. (2009). Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *J. Amer. Statist. Assoc.* **104** 572–585. [MR2751439](#) <https://doi.org/10.1198/jasa.2009.0023>
- CRAINICEANU, C. M. and GOLDSMITH, A. J. (2010). Bayesian functional data analysis using WinBUGS. *J. Stat. Softw.* **32** 1–33.
- CRAINICEANU, C. M., STAIU, A.-M. and DI, C.-Z. (2009). Generalized multilevel functional regression. *J. Amer. Statist. Assoc.* **104** 1550–1561. [MR2750578](#) <https://doi.org/10.1198/jasa.2009.tm08564>
- CURRIE, I. D., DURBAN, M. and EILERS, P. H. C. (2004). Smoothing and forecasting mortality rates. *Stat. Model.* **4** 279–298. [MR2086492](#) <https://doi.org/10.1191/1471082X04st080oa>
- HUMAN MORTALITY DATABASE (2021). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at <http://www.mortality.org> (data downloaded on 2018-12-24).
- DI, C.-Z., CRAINICEANU, C. M., CAFFO, B. S. and PUNJABI, N. M. (2009). Multilevel functional principal component analysis. *Ann. Appl. Stat.* **3** 458–488. [MR2668715](#) <https://doi.org/10.1214/08-AOAS206>
- GALLANT, A. R. (2009). *Nonlinear Statistical Models*. Wiley, Hoboken, NJ.
- GARCÍA-ESCUADERO, L. A. and GORDALIZA, A. (2005). A proposal for robust curve clustering. *J. Classification* **22** 185–201. [MR2231171](#) <https://doi.org/10.1007/s00357-005-0013-8>
- GIROSI, F. and KING, G. (2008). *Demographic Forecasting*. Princeton Univ. Press, Princeton, NJ.
- 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>
- GREVEN, S., CRAINICEANU, C., CAFFO, B. and REICH, D. (2011). Longitudinal functional principal component analysis. In *Recent Advances in Functional Data Analysis and Related Topics* (F. Ferraty, ed.). *Contrib. Statist.* 149–154. Physica-Verlag/Springer, Heidelberg. [MR2815575](#) https://doi.org/10.1007/978-3-7908-2736-1_23
- HALL, P. and VIAL, C. (2006). Assessing the finite dimensionality of functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 689–705. [MR2301015](#) <https://doi.org/10.1111/j.1467-9868.2006.00562.x>
- HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. [MR0666123](#) <https://doi.org/10.2307/1912775>
- HÖRMANN, S. and KIDZIŃSKI, Ł. (2015). A note on estimation in Hilbertian linear models. *Scand. J. Stat.* **42** 43–62. [MR3318024](#) <https://doi.org/10.1111/sjos.12094>
- HÖRMANN, S., KIDZIŃSKI, Ł. and HALLIN, M. (2015). Dynamic functional principal components. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 319–348. [MR3310529](#) <https://doi.org/10.1111/rssb.12076>
- HÖRMANN, S. and KOKOSZKA, P. (2012). Functional time series. In *Handbook of Statistics* (T. S. Rao, S. S. Rao and C. R. Rao, eds.) **30** 157–186. Elsevier, Amsterdam.
- HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. *Springer Series in Statistics*. Springer, New York. [MR2920735](#) <https://doi.org/10.1007/978-1-4614-3655-3>
- HORVÁTH, L., KOKOSZKA, P. and REEDER, R. (2013). Estimation of the mean of functional time series and a two-sample problem. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 103–122. [MR3008273](#) <https://doi.org/10.1111/j.1467-9868.2012.01032.x>
- HSIAO, C. (2014). *Analysis of Panel Data*, 3rd ed. *Econometric Society Monographs* **54**. Cambridge Univ. Press, New York. [MR3560738](#) <https://doi.org/10.1017/CBO9781139839327>
- HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.
- HYNDMAN, R. J. (2019). Demography: Forecasting mortality, fertility, migration and population data. Available at <https://CRAN.R-project.org/package=demography>, R package version 1.21.
- HYNDMAN, R. J. and ULLAH, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Comput. Statist. Data Anal.* **51** 4942–4956. [MR2364551](#) <https://doi.org/10.1016/j.csda.2006.07.028>
- HYNDMAN, R. J. and SHANG, H. L. (2009). Forecasting functional time series. *J. Korean Statist. Soc.* **38** 199–211. [MR2750314](#) <https://doi.org/10.1016/j.jkss.2009.06.002>
- HYNDMAN, R. J. and SHANG, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *J. Comput. Graph. Statist.* **19** 29–45. [MR2752026](#) <https://doi.org/10.1198/jcgs.2009.08158>
- HYNDMAN, R. J. and SHANG, H. L. (2021). *ftsa: Functional time series analysis*. Available at <https://CRAN.R-project.org/package=ftsar>, R package version 6.1.
- JACQUES, J. and PREDA, C. (2012). Model-based clustering of functional data. In *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges 459–464.
- JACQUES, J. and PREDA, C. (2014b). Functional data clustering: A survey. *Adv. Data Anal. Classif.* **8** 231–255. [MR3253859](#) <https://doi.org/10.1007/s11634-013-0158-y>

- JACQUES, J. and PRED, C. (2014c). 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](https://doi.org/10.1016/j.csda.2012.12.004)
- KARHUNEN, K. (1946). Zur Spektraltheorie stochastischer Prozesse. *Ann. Acad. Sci. Fenn., Ser. A I Math.-Phys.* **1946** 7. [MR0023012](https://doi.org/10.1016/j.csda.2012.12.004)
- KODINARIYA, T. M. and MAKWANA, P. R. (2013). Review on determining number of cluster in k-means clustering. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* **1** 90–95.
- LEE, R. D. and CARTER, L. R. (1992). Modeling and forecasting US mortality. *J. Amer. Statist. Assoc.* **87** 659–671.
- LI, J. (2013). A Poisson common factor model for projecting mortality and life expectancy jointly for females and males. *Popul. Stud.* **67** 111–126.
- LI, N. and LEE, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee–Carter method. *Demography* **42** 575–594.
- LI, N., LEE, R. and GERLAND, P. (2013). Extending the Lee–Carter method to model the rotation of age patterns of mortality decline for long-term projections. *Demography* **50** 2037–2051.
- LI, Y., WANG, N. and CARROLL, R. J. (2013). Selecting the number of principal components in functional data. *J. Amer. Statist. Assoc.* **108** 1284–1294. [MR3174708 https://doi.org/10.1080/01621459.2013.788980](https://doi.org/10.1080/01621459.2013.788980)
- LOÈVE, M. (1955). *Probability Theory. Foundations. Random Sequences*. D. Van Nostrand Co., Inc., Toronto–New York. [MR0066573](https://doi.org/10.1016/j.csda.2012.12.004)
- LÓPEZ-PINTADO, S. and ROMO, J. (2009). On the concept of depth for functional data. *J. Amer. Statist. Assoc.* **104** 718–734. [MR2541590 https://doi.org/10.1198/jasa.2009.0108](https://doi.org/10.1198/jasa.2009.0108)
- MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)* 281–297. Univ. California Press, Berkeley, CA. [MR0214227](https://doi.org/10.1016/j.csda.2012.12.004)
- MÜLLER, H.-G. (2005). Functional modelling and classification of longitudinal data. *Scand. J. Stat.* **32** 223–246. [MR2188671 https://doi.org/10.1111/j.1467-9469.2005.00429.x](https://doi.org/10.1111/j.1467-9469.2005.00429.x)
- NEWBY, W. K. and WEST, K. D. (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55** 703–708. [MR0890864 https://doi.org/10.2307/1913610](https://doi.org/10.2307/1913610)
- PAMPEL, F. (2005). Forecasting sex differences in mortality in high income nations: The contribution of smoking. *Demogr. Res.* **13** 455–484. <https://doi.org/10.4054/DemRes.2005.13.18>
- PANARETOS, V. M. and TAVAKOLI, S. (2013). Fourier analysis of stationary time series in function space. *Ann. Statist.* **41** 568–603. [MR3099114 https://doi.org/10.1214/13-AOS1086](https://doi.org/10.1214/13-AOS1086)
- POLITIS, D. N. and ROMANO, J. P. (1996). On flat-top kernel spectral density estimators for homogeneous random fields. *J. Statist. Plann. Inference* **51** 41–53. [MR1394143 https://doi.org/10.1016/0378-3758\(95\)00069-0](https://doi.org/10.1016/0378-3758(95)00069-0)
- POLITIS, D. N. and ROMANO, J. P. (1999). Multivariate density estimation with general flat-top kernels of infinite order. *J. Multivariate Anal.* **68** 1–25. [MR1668848 https://doi.org/10.1006/jmva.1998.1774](https://doi.org/10.1006/jmva.1998.1774)
- REISS, P. T. and OGDEN, R. T. (2007). Functional principal component regression and functional partial least squares. *J. Amer. Statist. Assoc.* **102** 984–996. [MR2411660 https://doi.org/10.1198/016214507000000527](https://doi.org/10.1198/016214507000000527)
- RENSHAW, A. E. and HABERMAN, S. (2003). Lee–Carter mortality forecasting with age-specific enhancement. *Insurance Math. Econom.* **33** 255–272. [MR2039286 https://doi.org/10.1016/S0167-6687\(03\)00138-0](https://doi.org/10.1016/S0167-6687(03)00138-0)
- RICE, G. and SHANG, H. L. (2017). A plug-in bandwidth selection procedure for long-run covariance estimation with stationary functional time series. *J. Time Series Anal.* **38** 591–609. [MR3664648 https://doi.org/10.1111/jtsa.12229](https://doi.org/10.1111/jtsa.12229)
- RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53** 233–243. [MR1094283](https://doi.org/10.1016/j.csda.2012.12.004)
- SERBAN, N. and JIANG, H. (2012). Multilevel functional clustering analysis. *Biometrics* **68** 805–814. [MR3055185 https://doi.org/10.1111/j.1541-0420.2011.01714.x](https://doi.org/10.1111/j.1541-0420.2011.01714.x)
- 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](https://doi.org/10.1198/016214504000001574)
- SHANG, H. L. (2016). Mortality and life expectancy forecasting for a group of populations in developed countries: A multilevel functional data method. *Ann. Appl. Stat.* **10** 1639–1672. [MR3553239 https://doi.org/10.1214/16-AOAS953](https://doi.org/10.1214/16-AOAS953)
- SHANG, H. L. (2018). Bootstrap methods for stationary functional time series. *Stat. Comput.* **28** 1–10. [MR3741633 https://doi.org/10.1007/s11222-016-9712-8](https://doi.org/10.1007/s11222-016-9712-8)
- SLIMEN, Y. B., ALLIO, S. and JACQUES, J. (2018). Model-based co-clustering for functional data. *Neurocomputing* **291** 97–108.
- SUGAR, C. A. and JAMES, G. M. (2003). Finding the number of clusters in a dataset: An information-theoretic approach. *J. Amer. Statist. Assoc.* **98** 750–763. [MR2012330 https://doi.org/10.1198/016214503000000666](https://doi.org/10.1198/016214503000000666)
- TANG, C., SHANG, H. L. and YANG, Y. (2022). Supplement to “Clustering and forecasting multiple functional time series.” <https://doi.org/10.1214/22-AOAS1602SUPPA>, <https://doi.org/10.1214/22-AOAS1602SUPPB>

- TARPEY, T. and KINATEDER, K. K. J. (2003). Clustering functional data. *J. Classification* **20** 93–114. [MR1983123 https://doi.org/10.1007/s00357-003-0007-3](https://doi.org/10.1007/s00357-003-0007-3)
- TSAY, R. S. and WOOD, D. (2021). MTS: All-purpose toolkit for analyzing multivariate time series and estimating multivariate volatility models. Available at <https://CRAN.R-project.org/package=MTS>, R package version 1.03.
- WARD, J. H. JR. (1963). Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.* **58** 236–244. [MR0148188](https://doi.org/10.1080/01621459.1963.10501372)
- WHITE, H. (1984). *Asymptotic Theory for Econometricians*. Academic Press, Cambridge, MA.
- WIŚNIEWSKI, A., SMITH, P. W., BIJAK, J., RAYMER, J. and FORSTER, J. J. (2015). Bayesian population forecasting: Extending the Lee–Carter method. *Demography* **52** 1035–1059.
- WOOLDRIDGE, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press, Cambridge, MA. [MR2768559](https://doi.org/10.1017/CBO9780511526461)
- YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. [MR2160561 https://doi.org/10.1198/016214504000001745](https://doi.org/10.1198/016214504000001745)

SEMI-SUPERVISED NONPARAMETRIC BAYESIAN MODELLING OF SPATIAL PROTEOMICS

BY OLIVER M. CROOK^{1,2,a}, KATHRYN S. LILLEY^{2,c}, LAURENT GATTO^{3,d} AND PAUL D. W. KIRK^{1,b}

¹MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, ^aoliver.crook@stats.ox.ac.uk,
^bpaul.kirk@mrc-bsu.cam.ac.uk

²Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, ^ck.s.lilley@bioc.cam.ac.uk

³de Duve Institute, UCLouvain, ^dlaurent.gatto@uclouvain.be

Understanding subcellular protein localisation is an essential component in the analysis of context specific protein function. Recent advances in quantitative mass-spectrometry (MS) have led to high-resolution mapping of thousands of proteins to subcellular locations within the cell. Novel modelling considerations to capture the complex nature of these data are thus necessary. We approach analysis of spatial proteomics data in a nonparametric Bayesian framework, using K-component mixtures of Gaussian process regression models. The Gaussian process regression model accounts for correlation structure within a subcellular niche, with each mixture component capturing the distinct correlation structure observed within each niche. The availability of *marker proteins* (i.e., proteins with a priori known labelled locations) motivates a semi-supervised learning approach to inform the Gaussian process hyperparameters. We moreover provide an efficient Hamiltonian-within-Gibbs sampler for our model. Furthermore, we reduce the computational burden associated with inversion of covariance matrices by exploiting the structure in the covariance matrix. A tensor decomposition of our covariance matrices allows extended Trench and Durbin algorithms to be applied to reduce the computational complexity of inversion and hence accelerate computation. We provide detailed case-studies on *Drosophila* embryos and mouse pluripotent embryonic stem cells to illustrate the benefit of semi-supervised functional Bayesian modelling of the data.

REFERENCES

- ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S. et al. (2000). Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25** 25–29.
- BARYLYUK, K., KORENY, L., KE, H., BUTTERWORTH, S., CROOK, O. M., LASSADI, I., GUPTA, V., TROMER, E. C., MOURIER, T. et al. (2020). A subcellular atlas of toxoplasma reveals the functional context of the proteome. *BioRxiv*.
- BELTRAN, P. M. J., MATHIAS, R. A. and CRISTEA, I. M. (2016). A portrait of the human organelle proteome in space and time during cytomegalovirus infection. *Cell Syst.* **3** 361–373.
- BESKOS, A., PILLAI, N., ROBERTS, G., SANZ-SERNA, J.-M. and STUART, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli* **19** 1501–1534. [MR3129023 https://doi.org/10.3150/12-BEJ414](https://doi.org/10.3150/12-BEJ414)
- BLOBEL, G. (2013). Christian de Duve (1917–2013). *Nature* **498** 300. <https://doi.org/10.1038/498300a>
- BOUVEYRON, C., CÔME, E. and JACQUES, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *Ann. Appl. Stat.* **9** 1726–1760. [MR3456352 https://doi.org/10.1214/15-AOAS861](https://doi.org/10.1214/15-AOAS861)
- BRECKELS, L. M., GATTO, L., CHRISTOFOROU, A., GROEN, A. J., LILLEY, K. S. and TROTTER, M. W. (2013). The effect of organelle discovery upon sub-cellular protein localisation. *J. Proteomics* **88** 129–140.
- BRECKELS, L. M., HOLDEN, S. B., WOJNAR, D., MULVEY, C. M., CHRISTOFOROU, A., GROEN, A., TROTTER, M. W., KOHLBACHER, O., LILLEY, K. S. et al. (2016). Learning from heterogeneous data sources: An application in spatial proteomics. *PLoS Comput. Biol.* **12** e1004920.

- CASELLA, G. and ROBERT, C. P. (1996). Rao-Blackwellisation of sampling schemes. *Biometrika* **83** 81–94. [MR1399157 https://doi.org/10.1093/biomet/83.1.81](https://doi.org/10.1093/biomet/83.1.81)
- CHRISTOFOROU, A., MULVEY, C. M., BRECKELS, L. M., GELADAKI, A., HURRELL, T., HAYWARD, P. C., NAAKE, T., GATTO, L., VINER, R. et al. (2016). A draft map of the mouse pluripotent stem cell spatial proteome. *Nat. Commun.* **7** 9992.
- CODY, N. A., IAMPIETRO, C. and LÉCUYER, E. (2013). The many functions of mRNA localization during normal development and disease: From pillar to post. *Wiley Interdiscip. Rev.: Dev. Biol.* **2** 781–796.
- COOK, K. C. and CRISTEA, I. M. (2019). Location is everything: Protein translocations as a viral infection strategy. *Curr Opin Chem Biol* **48** 34–43. <https://doi.org/10.1016/j.cbpa.2018.09.021>
- COOKE, E. J., SAVAGE, R. S., KIRK, P. D. W., DARKINS, R. and WILD, D. L. (2011). Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements. *BMC Bioinform.* **12** 399. <https://doi.org/10.1186/1471-2105-12-399>
- CORETTO, P. and HENNIG, C. (2016). Robust improper maximum likelihood: Tuning, computation, and a comparison with other methods for robust Gaussian clustering. *J. Amer. Statist. Assoc.* **111** 1648–1659. [MR3601724 https://doi.org/10.1080/01621459.2015.1100996](https://doi.org/10.1080/01621459.2015.1100996)
- CROOK, O. M., MULVEY, C. M., KIRK, P. D. W., LILLEY, K. S. and GATTO, L. (2018). A Bayesian mixture modelling approach for spatial proteomics. *PLoS Comput. Biol.* **14** 1–29.
- CROOK, O. M., GELADAKI, A., NIGHTINGALE, D. J. H., VENNARD, O. L., LILLEY, K. S., GATTO, L. and KIRK, P. D. W. (2020). A semi-supervised Bayesian approach for simultaneous protein sub-cellular localisation assignment and novelty detection. *PLoS Comput. Biol.* **16** e1008288. <https://doi.org/10.1371/journal.pcbi.1008288>
- CROOK, O. M., LILLEY, K. S., GATTO, L. and KIRK, P. D. (2022). Supplement to “Semi-Supervised Nonparametric Bayesian Modelling of Spatial Proteomics.” <https://doi.org/10.1214/22-AOAS1603SUPP>
- DAVIES, A. K., ITZHAK, D. N., EDGAR, J. R., ARCHULETA, T. L., HIRST, J., JACKSON, L. P., ROBINSON, M. S. and BORNER, G. H. H. (2018). AP-4 vesicles contribute to spatial control of autophagy via RUSC-dependent peripheral delivery of ATG9A. *Nat. Commun.* **9** 3958. <https://doi.org/10.1038/s41467-018-06172-7>
- DE DUVE, C. (1969). The peroxisome: A new cytoplasmic organelle. *Proc. R. Soc. Lond., B Biol. Sci.* **173** 71–83.
- DE DUVE, C. and BEAUFAY, H. (1981). A short history of tissue fractionation. *J. Cell Biol.* **91** 293.
- DE MATTEIS, M. A. and LUINI, A. (2011). Mendelian disorders of membrane trafficking. *N. Engl. J. Med.* **365** 927–938.
- DUANE, S., KENNEDY, A. D., PENDLETON, B. J. and ROWETH, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195** 216–222. [MR3960671 https://doi.org/10.1016/0370-2693\(87\)91197-x](https://doi.org/10.1016/0370-2693(87)91197-x)
- DUNKLEY, T. P., WATSON, R., GRIFFIN, J. L., DUPREE, P. and LILLEY, K. S. (2004). Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics* **3** 1128–1134.
- DUNKLEY, T. P., HESTER, S., SHADFORTH, I. P., RUNIONS, J., WEIMAR, T., HANTON, S. L., GRIFFIN, J. L., BESSANT, C., BRANDIZZI, F. et al. (2006). Mapping the arabidopsis organelle proteome. *Proc. Natl. Acad. Sci. USA* **103** 6518–6523.
- EDDELBUEITTEL, D. and FRANCOIS, R. (2011). Rcpp: Seamless R and C++ integration. *J. Stat. Softw.* **40** 1–18.
- EDDELBUEITTEL, D. and SANDERSON, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Comput. Statist. Data Anal.* **71** 1054–1063. [MR3132026 https://doi.org/10.1016/j.csda.2013.02.005](https://doi.org/10.1016/j.csda.2013.02.005)
- ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. [MR1340510](https://doi.org/10.1080/01621459.2015.1100996)
- FRALEY, C. and RAFTERY, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *J. Classification* **24** 155–181. [MR2415725 https://doi.org/10.1007/s00357-007-0004-5](https://doi.org/10.1007/s00357-007-0004-5)
- GATTO, L., BRECKELS, L. M. and LILLEY, K. S. (2019). Assessing sub-cellular resolution in spatial proteomics experiments. *Curr. Opin. Chem. Biol.* **48** 123–149. <https://doi.org/10.1016/j.cbpa.2018.11.015>
- GATTO, L., CROOK, O. M. and BRECKELS, L. M. (2018). pRolocdata: Data accompanying the pRoloc package. R package version 1.19.1.
- GATTO, L., VIZCAÍNO, J. A., HERMIAKOB, H., HUBER, W. and LILLEY, K. S. (2010). Organelle proteomics experimental designs and analysis. *Proteomics* **10** 3957–3969.
- GATTO, L., BRECKELS, L. M., BURGER, T., NIGHTINGALE, D. J., GROEN, A. J., CAMPBELL, C., MULVEY, C. M., CHRISTOFOROU, A., FERRO, M. et al. (2014a). A foundation for reliable spatial proteomics data analysis. *Mol. Cell. Proteomics* mcp–M113.
- GATTO, L., BRECKELS, L. M., WIECZOREK, S., BURGER, T. and LILLEY, K. S. (2014b). Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics* **30** 1322–1324. <https://doi.org/10.1093/bioinformatics/btu013>
- GELADAKI, A., BRITOVSEK, N. K., BRECKELS, L. M., SMITH, T. S. O. L. V., MULVEY, C. M., CROOK, O. M., GATTO, L. and LILLEY, K. S. (2019). Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nat. Commun.* **10** 331.

- GELFAND, A. E., KOTTAS, A. and MACEACHERN, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100** 1021–1035. [MR2201028](#) <https://doi.org/10.1198/016214504000002078>
- GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. [MR1141740](#)
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. [MR2221284](#) <https://doi.org/10.1214/06-BA117A>
- GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (1995). *Bayesian Data Analysis. Texts in Statistical Science Series*. CRC Press, London. [MR1385925](#)
- GENTLEMAN, R. C., CAREY, V. J., BATES, D. M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y. et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **5** R80.
- GIBSON, T. J. (2009). Cell regulation: Determined to signal discrete cooperation. *Trends Biochem. Sci.* **34** 471–482.
- 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>
- GROEN, A. J., SANCHO-ANDRÉ, G., BRECKELS, L. M., GATTO, L., ANIENTO, F. and LILLEY, K. S. (2014). Identification of trans-Golgi network proteins in Arabidopsis thaliana root tissue. *J. Proteome Res.* **13** 763–776.
- HALL, S. L., HESTER, S., GRIFFIN, J. L., LILLEY, K. S. and JACKSON, A. P. (2009). The organelle proteome of the DT40 lymphocyte cell line. *Mol. Cell. Proteomics* **8** 1295–1305.
- HEARD, W., SKLENÁŘ, J., TOME, D. F., ROBATZEK, S. and JONES, A. M. (2015). Identification of regulatory and cargo proteins of endosomal and secretory pathways in arabidopsis thaliana by proteomic dissection. *Mol. Cell. Proteomics* **14** 1796–1813.
- HEINONEN, M., GUIPAUD, O., MILLIAT, F., BUARD, V., MICHEAU, B., TARLET, G., BENDERITTER, M., ZEHRAOUI, F. and D'ALCHÉ BUC, F. (2014). Detecting time periods of differential gene expression using Gaussian processes: An application to endothelial cells exposed to radiotherapy dose fraction. *Bioinformatics* **31** 728–735.
- HENNIG, C. (2004). Breakdown points for maximum likelihood estimators of location-scale mixtures. *Ann. Statist.* **32** 1313–1340. [MR2089126](#) <https://doi.org/10.1214/009053604000000571>
- HENSMAN, J., LAWRENCE, N. D. and RATTRAY, M. (2013). Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinform.* **14** 252. <https://doi.org/10.1186/1471-2105-14-252>
- HIRST, J., ITZHAK, D. N., ANTROBUS, R., BORNER, G. H. H. and ROBINSON, M. S. (2018). Role of the AP-5 adaptor protein complex in late endosome-to-Golgi retrieval. *PLoS Biol.* **16** e2004411. <https://doi.org/10.1371/journal.pbio.2004411>
- HONKELA, A., GIRARDOT, C., GUSTAFSON, E. H., LIU, Y.-H., FURLONG, E. E., LAWRENCE, N. D. and RATTRAY, M. (2010). Model-based method for transcription factor target identification with limited data. *Proc. Natl. Acad. Sci. USA* **107** 7793–7798.
- HUBER, W., CAREY, V. J., GENTLEMAN, R., ANDERS, S., CARLSON, M., CARVALHO, B. S., BRAVO, H. C., DAVIS, S., GATTO, L. et al. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nat. Methods* **12** 115.
- ITZHAK, D. N., TYANOVA, S., COX, J. and BORNER, G. H. (2016). Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* **5**. <https://doi.org/10.7554/eLife.16950>
- ITZHAK, D. N., DAVIES, C., TYANOVA, S., MISHRA, A., WILLIAMSON, J., ANTROBUS, R., COX, J., WEEKES, M. P. and BORNER, G. H. H. (2017). A mass spectrometry-based approach for mapping protein subcellular localization reveals the spatial proteome of mouse primary neurons. *Cell Rep.* **20** 2706–2718. <https://doi.org/10.1016/j.celrep.2017.08.063>
- JADOT, M., BOONEN, M., THIRION, J., WANG, N., XING, J., ZHAO, C., TANNOUN, A., QIAN, M., ZHENG, H. et al. (2017). Accounting for protein subcellular localization: A compartmental map of the rat liver proteome. *Mol. Cell. Proteomics* **16** 194–212.
- JAMES, G. M. and HASTIE, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 533–550. [MR1858401](#) <https://doi.org/10.1111/1467-9868.00297>
- 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>
- JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A* **186** 453–461. [MR0017504](#) <https://doi.org/10.1098/rspa.1946.0056>
- JONES, M. and RICE, J. A. (1992). Displaying the important features of large collections of similar curves. *Amer. Statist.* **46** 140–145.

- KALAITZIS, A. A. and LAWRENCE, N. D. (2011a). A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinform.* **12** 180. <https://doi.org/10.1186/1471-2105-12-180>
- KALAITZIS, A. A. and LAWRENCE, N. D. (2011b). A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinform.* **12** 180. <https://doi.org/10.1186/1471-2105-12-180>
- KAU, T. R., WAY, J. C. and SILVER, P. A. (2004). Nuclear transport and cancer: From mechanism to intervention. *Nat. Rev. Cancer* **4** 106–117.
- KIRK, P. D. and STUMPF, M. P. (2009). Gaussian process regression bootstrapping: Exploring the effects of uncertainty in time course data. *Bioinformatics* **25** 1300–1306.
- KIRK, P., GRIFFIN, J. E., SAVAGE, R. S., GHAHRAMANI, Z. and WILD, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28** 3290–3297.
- LATORRE, I. J., ROH, M. H., FRESE, K. K., WEISS, R. S., MARGOLIS, B. and JAVIER, R. T. (2005). Viral oncoprotein-induced mislocalization of select PDZ proteins disrupts tight junctions and causes polarity defects in epithelial cells. *J. Cell Sci.* **118** 4283–4293.
- LAURILA, K. and VIHINEN, M. (2009). Prediction of disease-related mutations affecting protein localization. *BMC Genomics* **10** 122. <https://doi.org/10.1186/1471-2164-10-122>
- LAVINE, M. and WEST, M. (1992). A Bayesian method for classification and discrimination. *Canad. J. Statist.* **20** 451–461. [MR1208356 https://doi.org/10.2307/3315614](https://doi.org/10.2307/3315614)
- LIU, D. C. and NOCEDAL, J. (1989). On the limited memory BFGS method for large scale optimization. *Math. Program.* **45** 503–528. [MR1038245 https://doi.org/10.1007/BF01589116](https://doi.org/10.1007/BF01589116)
- LIU, Q., LIN, K. K., ANDERSEN, B., SMYTH, P. and IHLER, A. (2010). Estimating replicate time shifts using Gaussian process regression. *Bioinformatics* **26** 770–776.
- LUHESHI, L. M., CROWTHER, D. C. and DOBSON, C. M. (2008). Protein misfolding and disease: From the test tube to the organism. *Curr. Opin. Chem. Biol.* **12** 25–31.
- LUNN, D. J., THOMAS, A., BEST, N. and SPIEGELHALTER, D. (2000). WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Stat. Comput.* **10** 325–337.
- MACHETE, R. L. (2013). Contrasting probabilistic scoring rules. *J. Statist. Plann. Inference* **143** 1781–1790. [MR3082233 https://doi.org/10.1016/j.jspi.2013.05.012](https://doi.org/10.1016/j.jspi.2013.05.012)
- MALSINER-WALLI, G., FRÜHWIRTH-SCHNATTER, S. and GRÜN, B. (2017). Identifying mixtures of mixtures using Bayesian estimation. *J. Comput. Graph. Statist.* **26** 285–295. [MR3640186 https://doi.org/10.1080/10618600.2016.1200472](https://doi.org/10.1080/10618600.2016.1200472)
- MENDES, M., PELÁEZ-GARCÍA, A., LÓPEZ-LUCENDO, M., BARTOLOMÉ, R. A., CALVIÑO, E., BARDERAS, R. and CASAL, J. I. (2017). Mapping the spatial proteome of metastatic cells in colorectal cancer. *Proteomics* **17**. <https://doi.org/10.1002/pmic.201700094>
- METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.
- MORRIS, J. S. (2015). Functional regression. *Annu. Rev. Stat. Appl.* **2** 321–359.
- MULVEY, C. M., BRECKELS, L. M., GELADAKI, A., BRITOVŠEK, N. K., NIGHTINGALE, D. J. H., CHRISTOFOROU, A., ELZEK, M., DEERY, M. J., GATTO, L. et al. (2017). Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. *Nat. Protoc.* **12** 1110–1135. <https://doi.org/10.1038/nprot.2017.026>
- MURPHY, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge.
- MURPHY, K. and MURPHY, T. B. (2019). Parsimonious model-based clustering with covariates. *Adv. Data Anal. Classif.*
- NIGHTINGALE, D. J., GELADAKI, A., BRECKELS, L. M., OLIVER, S. G. and LILLEY, K. S. (2019). The subcellular organisation of *saccharomyces cerevisiae*. *Curr. Opin. Chem. Biol.* **48** 86–95. <https://doi.org/10.1016/j.cbpa.2018.10.026>
- NIKOLOVSKI, N., RUBTSOV, D., SEGURA, M. P., MILES, G. P., STEVENS, T. J., DUNKLEY, T. P., MUNRO, S., LILLEY, K. S. and DUPREE, P. (2012). Putative glycosyltransferases and other plant Golgi apparatus proteins are revealed by LOPIT proteomics. *Plant Physiol.* **160** 1037–1051.
- OHTA, S., BUKOWSKI-WILLS, J.-C., SANCHEZ-PULIDO, L., DE LIMA ALVES, F., WOOD, L., CHEN, Z. A., PLATANI, M., FISCHER, L., HUDSON, D. F. et al. (2010). The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics. *Cell* **142** 810–821.
- OLKKONEN, V. M. and IKONEN, E. (2006). When intracellular logistics fails-genetic defects in membrane trafficking. *J. Cell Sci.* **119** 5031–5045.
- ORRE, L. M., VESTERLUND, M., PAN, Y., ARSLAN, T., ZHU, Y., WOODBRIDGE, A. F., FRINGS, O., FREDLUND, E. and LEHTIÖ, J. (2019). SubCellBarCode: Proteome-wide mapping of protein localization and relocalization. *Mol. Cell* **73** 166–182.e7.

- PARSONS, H., FERNÁNDEZ-NIÑO, S. and HEAZLEWOOD, J. (2014). Separation of the plant Golgi apparatus and endoplasmic reticulum by free-flow electrophoresis. *Methods Mol. Biol. (Clifton N.J.)* **1072** 527.
- PREDÁ, C., SAPORTA, G. and LÉVÉDER, C. (2007). PLS classification of functional data. *Comput. Statist.* **22** 223–235. [MR2318457 https://doi.org/10.1007/s00180-007-0041-4](https://doi.org/10.1007/s00180-007-0041-4)
- RAMSAY, J. O. (2004). Functional data analysis. *Encyc. Stat. Sci.* **4**.
- RASMUSSEN, C. E. (2004). Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning* 63–71. Springer, Berlin.
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR2514435](https://doi.org/10.1017/9780521876223)
- RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792. [MR1483213 https://doi.org/10.1111/1467-9868.00095](https://doi.org/10.1111/1467-9868.00095)
- RODRIGUEZ, J. A., AU, W. W. Y. and HENDERSON, B. R. (2004). Cytoplasmic mislocalization of BRCA1 caused by cancer-associated mutations in the BRCT domain. *Exp. Cell Res.* **293** 14–21. <https://doi.org/10.1016/j.yexcr.2003.09.027>
- RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2009). Bayesian nonparametric functional data analysis through density estimation. *Biometrika* **96** 149–162. [MR2482141 https://doi.org/10.1093/biomet/asn054](https://doi.org/10.1093/biomet/asn054)
- SADOWSKI, P. G., DUNKLEY, T. P., SHADFORTH, I. P., DUPREE, P., BESSANT, C., GRIFFIN, J. L. and LILLEY, K. S. (2006). Quantitative proteomic approach to study subcellular localization of membrane proteins. *Nat. Protoc.* **1** 1778–1789.
- SCHUURMAN, N. K., GRASMAN, R. P. P. and HAMAKER, E. L. A comparison of inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivar. Behav. Res.* **51** 185–206. <https://doi.org/10.1080/00273171.2015.1065398>
- SHANNON, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27** 379–423. [MR0026286 https://doi.org/10.1002/j.1538-7305.1948.tb01338.x](https://doi.org/10.1002/j.1538-7305.1948.tb01338.x)
- SHIN, S. J., SMITH, J. A., REZNICZEK, G. A., PAN, S., CHEN, R., BRETNALL, T. A., WICHE, G. and KELLY, K. A. (2013). Unexpected gain of function for the scaffolding protein plectin due to mislocalization in pancreatic cancer. *Proc. Natl. Acad. Sci. USA* **110** 19414–19419.
- SHIN, J. J., CROOK, O. M., BORGEAUD, A., CATTIN-ORTOLÁ, J., PEAK-CHEW, S.-Y., CHADWICK, J., LILLEY, K. S. and MUNRO, S. (2019). Determining the content of vesicles captured by golgin tethers using LOPIT-DC. *BioRxiv* 841965.
- SILJEE, J. E., WANG, Y., BERNARD, A. A., ERSOY, B. A., ZHANG, S., MARLEY, A., VON ZASTROW, M., REITER, J. F. and VAISSE, C. (2018). Subcellular localization of MC4R with ADCY3 at neuronal primary cilia underlies a common pathway for genetic predisposition to obesity. *Nat. Genet.*
- STEEL, M. F. and FUENTES, M. (2010). *Non-Gaussian and Nonparametric Models for Continuous Spatial Data*. CRC Press, Boca Raton, FL.
- STEGLE, O., DENBY, K. J., COOKE, E. J., WILD, D. L., GHAMRANI, Z. and BORGWARDT, K. M. (2010). A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *J. Comput. Biol.* **17** 355–367. [MR2609139 https://doi.org/10.1089/cmb.2009.0175](https://doi.org/10.1089/cmb.2009.0175)
- TAN, D. J., DVINGE, H., CHRISTOFOROU, A., BERTONE, P., MARTINEZ ARIAS, A. and LILLEY, K. S. (2009). Mapping organelle proteins and protein complexes in drosophila melanogaster. *J. Proteome Res.* **8** 2667–2678.
- TARDIF, M., ATTEIA, A., SPECHT, M., COGNE, G., ROLLAND, N., BRUGIÈRE, S., HIPPLER, M., FERRO, M., BRULEY, C. et al. (2012). PredAlgo: A new subcellular localization prediction tool dedicated to green algae. *Mol. Biol. Evol.* **29** 3625–3639.
- THOMPSON, A., SCHÄFER, J., KUHN, K., KIENLE, S., SCHWARZ, J., SCHMIDT, G., NEUMANN, T. and HAMON, C. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75** 1895–1904.
- THUL, P. J., ÅKESSON, L., WIKING, M., MAHDESSIAN, D., GELADAKI, A., AIT BLAL, H., ALM, T., ASPLUND, A., BJÖRK, L. et al. (2017). A subcellular map of the human proteome. *Science*.
- TOPA, H., JÓNÁS, Á., KOFLER, R., KOSIOL, C. and HONKELA, A. (2015). Gaussian process test for high-throughput sequencing time series: Application to experimental evolution. *Bioinformatics* **31** 1762–1770.
- WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.
- WANG, L. and DUNSON, D. B. (2011). Fast Bayesian inference in Dirichlet process mixture models. *J. Comput. Graph. Statist.* **20** 196–216. Supplementary material available online. [MR2816545 https://doi.org/10.1198/jcgs.2010.07081](https://doi.org/10.1198/jcgs.2010.07081)
- WILLIAMS, C. K. and RASMUSSEN, C. E. (1996). Gaussian processes for regression. In *Advances in Neural Information Processing Systems* 514–520.
- ZHANG, Y., LEITHEAD, W. E. and LEITH, D. J. (2005). Time-series Gaussian process regression based on Toeplitz computation of $O(N^2)$ operations and $O(N)$ -level storage. In *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference on* 3711–3716. IEEE, Los Alamitos.

- ZHU, H., BROWN, P. J. and MORRIS, J. S. (2012). Robust classification of functional and quantitative image data using functional mixed models. *Biometrics* **68** 1260–1268. [MR3040032](#) <https://doi.org/10.1111/j.1541-0420.2012.01765.x>
- ZHU, H., VANNUCCI, M. and COX, D. D. (2010). A Bayesian hierarchical model for classification with selection of functional predictors. *Biometrics* **66** 463–473. [MR2758826](#) <https://doi.org/10.1111/j.1541-0420.2009.01283.x>

BAYESIAN DATA SYNTHESIS AND THE UTILITY-RISK TRADE-OFF FOR MIXED EPIDEMIOLOGICAL DATA

BY JOSEPH FELDMAN^a AND DANIEL R. KOWAL^b

Department of Statistics, Rice University, ^ajoseph.r.feldman@rice.edu, ^bdaniel.kowal@rice.edu

Much of the microdata used for epidemiological studies contain sensitive measurements on real individuals. As a result, such microdata cannot be published out of privacy concerns, and without public access to these data, any statistical analyses originally published on them are nearly impossible to reproduce. To promote the dissemination of key datasets for analysis without jeopardizing the privacy of individuals, we introduce a cohesive Bayesian framework for the generation of fully synthetic high-dimensional microdatasets of mixed categorical, binary, count, and continuous variables. This process centers around a joint Bayesian model that is simultaneously compatible with all of these data types, enabling the creation of mixed synthetic datasets through posterior predictive sampling. Furthermore, a focal point of epidemiological data analysis is the study of conditional relationships between various exposures and key outcome variables through regression analysis. We design a modified data synthesis strategy to target and preserve these conditional relationships, including both nonlinearities and interactions. The proposed techniques are deployed to create a synthetic version of a confidential dataset containing dozens of health, cognitive, and social measurements on nearly 20,000 North Carolina children.

REFERENCES

- ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. [MR1224394](#)
- BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. [MR2806429](#) <https://doi.org/10.1093/biomet/asr013>
- CAIOLA, G. and REITER, J. P. (2010). Random forests for generating partially synthetic, categorical data. *Trans. Data Priv.* **3** 27–42. [MR2725418](#)
- 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>
- 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>
- DRECHSLER, J. (2018). Some clarifications regarding fully synthetic data. In *International Conference on Privacy in Statistical Databases* 109–121. Springer, Berlin.
- DUNCAN, G. T., KELLER-MCNULTY, S. A. and STOKES, S. L. (2001). Disclosure risk vs data utility: The R-U confidentiality map.
- 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>
- ELLIOT, M. (2015). Final report on the disclosure risk associated with the synthetic data produced by the sylls team. *Report 2015* **2**.
- FELDMAN, J. and KOWAL, D. R. (2022). Supplement to “Bayesian data synthesis and the utility-risk trade-off for mixed epidemiological data.” <https://doi.org/10.1214/22-AOAS1604SUPPA>, <https://doi.org/10.1214/22-AOAS1604SUPPB>
- HOFF, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.* **1** 265–283. [MR2393851](#) <https://doi.org/10.1214/07-AOAS107>
- HU, J. (2019). Bayesian estimation of attribute and identification disclosure risks in synthetic data. 61–89.
- HU, J., REITER, J. P. and WANG, Q. (2014). Disclosure risk evaluation for fully synthetic categorical data. In *International Conference on Privacy in Statistical Databases* 185–199. Springer, Berlin.

- JOHNDROW, J., DUNSON, D. and LUM, K. (2013). Diagonal orthant multinomial probit models. In *Artificial Intelligence and Statistics* 29–38.
- KINNEY, S. K., REITER, J. P., REZNEK, A. P., MIRANDA, J., JARMIN, R. S. and ABOWD, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *Int. Stat. Rev.* **79** 362–384.
- KOWAL, D. R. (2021). Fast, optimal, and targeted predictions using parameterized decision analysis. *J. Amer. Statist. Assoc.* 1–12.
- KOWAL, D. R. and CANALE, A. (2020). Simultaneous transformation and rounding (STAR) models for integer-valued data. *Electron. J. Stat.* **14** 1744–1772. [MR4083734](#) <https://doi.org/10.1214/20-EJS1707>
- KOWAL, D. R., BRAVO, M., LEONG, H., BUI, A., GRIFFIN, R. J., ENSOR, K. B. and MIRANDA, M. L. (2021). Bayesian variable selection for understanding mixtures in environmental exposures. *Stat. Med.* **40** 4850–4871. [MR4315456](#) <https://doi.org/10.1002/sim.9099>
- LITTLE, R. J. (1993). Statistical analysis of masked data. *J. Off. Stat.* **9** 407.
- MIRANDA, M. L., KIM, D., GALEANO, M. A. O., PAUL, C. J., HULL, A. P. and MORGAN, S. P. (2007). The relationship between early childhood blood lead levels and performance on end-of-grade tests. *Environ. Health Perspect.* **115** 1242–1247.
- MURRAY, J. S. and REITER, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *J. Amer. Statist. Assoc.* **111** 1466–1479. [MR3601702](#) <https://doi.org/10.1080/01621459.2016.1174132>
- MURRAY, J. S., DUNSON, D. B., CARIN, L. and LUCAS, J. E. (2013). Bayesian Gaussian copula factor models for mixed data. *J. Amer. Statist. Assoc.* **108** 656–665. [MR3174649](#) <https://doi.org/10.1080/01621459.2012.762328>
- NOWOK, B., RAAB, G., SNOKE, J. and DIBBEN, C. (2016). synthpop: Generating synthetic versions of sensitive microdata for statistical disclosure control. *R Package Version* 1–3.
- QUICK, H., HOLAN, S. H., WIKLE, C. K. and REITER, J. P. (2015). Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography. *Spat. Stat.* **14** 439–451. [MR3431050](#) <https://doi.org/10.1016/j.spasta.2015.07.008>
- RAAB, G. M., NOWOK, B. and DIBBEN, C. (2016). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality* **7** 67–97.
- REITER, J. P. (2005a). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29** 181–188.
- REITER, J. P. (2005b). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *J. Roy. Statist. Soc. Ser. A* **168** 185–205. [MR2113234](#) <https://doi.org/10.1111/j.1467-985X.2004.00343.x>
- REITER, J. P. (2005c). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics* **21** 441.
- REITER, J. P. and MITRA, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality* **1**.
- REITER, J. P., WANG, Q. and ZHANG, B. (2014). Bayesian estimation of disclosure risks for multiply imputed, synthetic data. *Journal of Privacy and Confidentiality* **6**.
- RUBIN, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics* **9** 461–468.
- SKLAR, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8** 229–231. [MR0125600](#)
- SNOKE, J., RAAB, G. M., NOWOK, B., DIBBEN, C. and SLAVKOVIC, A. (2018). General and specific utility measures for synthetic data. *J. Roy. Statist. Soc. Ser. A* **181** 663–688. [MR3807503](#) <https://doi.org/10.1111/rssa.12358>
- TAUB, J. (2021). Synthetic data: An exploration of data utility and disclosure risk. 138–160.
- TAUB, J., ELLIOT, M., PAMPAKA, M. and SMITH, D. (2018). Differential correct attribution probability for synthetic data: An exploration. In *International Conference on Privacy in Statistical Databases* 122–137. Springer, Berlin.

HIERARCHICAL BAYESIAN MODELING OF OCEAN HEAT CONTENT AND ITS UNCERTAINTY

BY SAMUEL BAUGH^{1,a} AND KAREN MCKINNON^{2,b}

¹Department of Statistics, UCLA, ^asamuelbaugh@ucla.edu

²Department of Statistics, UCLA and Institute for the Environment and Sustainability, UCLA, ^bkmckinnon@ucla.edu

The accurate quantification of changes in the heat content of the world's oceans is crucial for our understanding of the effects of increasing greenhouse gas concentrations. The Argo program, consisting of Lagrangian floats that measure vertical temperature profiles throughout the global ocean, has provided a wealth of data from which to estimate ocean heat content. However, creating a globally consistent statistical model for ocean heat content remains challenging due to the need for a globally valid covariance model that can capture complex nonstationarity. In this paper, we develop a hierarchical Bayesian Gaussian process model that uses kernel convolutions with cylindrical distances to allow for spatial nonstationarity in all model parameters while using a Vecchia process to remain computationally feasible for large spatial datasets. Our approach can produce valid credible intervals for globally integrated quantities that would not be possible using previous approaches. These advantages are demonstrated through the application of the model to Argo data, yielding credible intervals for the spatially varying trend in ocean heat content that accounts for both the uncertainty induced from interpolation and from estimating the mean field and other parameters. Through cross-validation, we show that our model outperforms an out-of-the-box approach as well as other simpler models. The code for performing this analysis is provided as the R package **BayesianOHC**.

REFERENCES

- BALMASEDA, M. A., TRENBERTH, K. E. and KÄLLÉN, E. (2013). Distinctive climate signals in reanalysis of global ocean heat content. *Geophys. Res. Lett.* **40** 1754–1759.
- BAUGH, S. and MCKINNON, K. (2022). Supplement to “Hierarchical Bayesian modeling of ocean heat content and its uncertainty.” <https://doi.org/10.1214/22-AOAS1605SUPPA>, <https://doi.org/10.1214/22-AOAS1605SUPPB>
- CHENG, L. and ZHU, J. (2016). Benefits of CMIP5 multimodel ensemble in reconstructing historical ocean subsurface temperature variations. *J. Climate* **29** 5393–5416.
- CHENG, L., TRENBERTH, K. E., FASULLO, J., BOYER, T., ABRAHAM, J. and ZHU, J. (2017). Improved estimates of ocean heat content from 1960 to 2015. *Sci. Adv.* **3**.
- COLLING, A. (2001). *Ocean Circulation* **3**. Butterworth-Heinemann Boston.
- DATA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. [MR3538706 https://doi.org/10.1080/01621459.2015.1044091](https://doi.org/10.1080/01621459.2015.1044091)
- DIJKSTRA, H. A. (2008). *Dynamical Oceanography*. Springer, Berlin. [MR2723932 https://doi.org/10.1007/978-3-540-77777-7](https://doi.org/10.1007/978-3-540-77777-7)
- GNEITING, T., RAFTERY, A. E., WESTVELD III, A. H. and GOLDMAN, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133** 1098–1118.
- GOURETSKI, V. (2018). World ocean circulation experiment—Argo global hydrographic climatology. *Ocean Science* **14** 1127–1146.
- GUINNESS, J. (2018). Permutation and grouping methods for sharpening Gaussian process approximations. *Technometrics* **60** 415–429. [MR3878098 https://doi.org/10.1080/00401706.2018.1437476](https://doi.org/10.1080/00401706.2018.1437476)
- GUINNESS, J. (2021). Gaussian process learning via Fisher scoring of Vecchia's approximation. *Stat. Comput.* **31** Paper No. 25. [MR4224950 https://doi.org/10.1007/s11222-021-09999-1](https://doi.org/10.1007/s11222-021-09999-1)

- GUINNESS, J. and GPGP, M. K. (2018). Fast Gaussian process computation using Vecchia's approximation. R package version 0.1.0.
- HEIDELBERGER, P. and WELCH, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Commun. ACM* **24** 233–245. [MR0611745 https://doi.org/10.1145/358598.358630](https://doi.org/10.1145/358598.358630)
- HIGDON, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environ. Ecol. Stat.* **5** 173–190.
- HITCZENKO, M. and STEIN, M. L. (2012). Some theory for anisotropic processes on the sphere. *Stat. Methodol.* **9** 211–227. [MR2863609 https://doi.org/10.1016/j.stamet.2011.01.010](https://doi.org/10.1016/j.stamet.2011.01.010)
- ISHII, M., FUKUDA, Y., HIRAHARA, S., YASUI, S., SUZUKI, T. and SATO, K. (2017). Accuracy of global upper ocean heat content estimation expected from present observational data sets. *Sola* **13** 163–167.
- JEONG, J. and JUN, M. (2015). A class of Matérn-like covariance functions for smooth processes on a sphere. *Spat. Stat.* **11** 1–18. [MR3311853 https://doi.org/10.1016/j.spasta.2014.11.001](https://doi.org/10.1016/j.spasta.2014.11.001)
- JEONG, J., JUN, M. and GENTON, M. G. (2017). Spherical process models for global spatial statistics. *Statist. Sci.* **32** 501–513. [MR3730519 https://doi.org/10.1214/17-STS620](https://doi.org/10.1214/17-STS620)
- KATZFUSS, M. (2013). Bayesian nonstationary spatial modeling for very large datasets. *Environmetrics* **24** 189–200. [MR3067342 https://doi.org/10.1002/env.2200](https://doi.org/10.1002/env.2200)
- KATZFUSS, M. and GUINNESS, J. (2021). A general framework for Vecchia approximations of Gaussian processes. *Statist. Sci.* **36** 124–141. [MR4194207 https://doi.org/10.1214/19-STS755](https://doi.org/10.1214/19-STS755)
- KATZFUSS, M., JUREK, M., ZILBER, D., GONG, W., GUINNESS, J., ZHANG, J. and GPVECCHIA, F. S. (2020a). Fast Gaussian-process inference using Vecchia approximations. R package version 0.1.
- KATZFUSS, M., GUINNESS, J., GONG, W. and ZILBER, D. (2020b). Vecchia approximations of Gaussian-process predictions. *J. Agric. Biol. Environ. Stat.* **25** 383–414. [MR4139037 https://doi.org/10.1007/s13253-020-00401-7](https://doi.org/10.1007/s13253-020-00401-7)
- KUUSELA, M. and STEIN, M. L. (2018). Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **474** 20180400.
- LEVITUS, S., ANTONOV, J. I., BOYER, T. P., BARANOVA, O. K., GARCIA, H. E., LOCARNINI, R. A., MISHONOV, A. V., REAGAN, J. R., SEIDOV, D. et al. (2012). World ocean heat content and thermosteric sea level change (0–2000 m), 1955–2010. *Geophysical Research Letters* **39** L10603.
- LI, Y. and ZHU, Z. (2016). Modeling nonstationary covariance function with convolution on sphere. *Comput. Statist. Data Anal.* **104** 233–246. [MR3540998 https://doi.org/10.1016/j.csda.2016.07.001](https://doi.org/10.1016/j.csda.2016.07.001)
- 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](https://doi.org/10.1111/j.1467-9868.2011.00777.x)
- MEYSSIGNAC, B., BOYER, T., ZHAO, Z., HAKUBA, M. Z., LANDERER, F. W., STAMMER, D., KÖHL, A., KATO, S., L'ECUYER, T. et al. (2019). Measuring global ocean heat content to estimate the Earth energy imbalance. *Front. Mar. Sci.* **6** 432.
- PACIOREK, C. J. and SCHERVISH, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* **17** 483–506. [MR2240939 https://doi.org/10.1002/env.785](https://doi.org/10.1002/env.785)
- PALMER, M. D., ROBERTS, C. D., BALMASEDA, M. A. et al. (2017). Ocean heat content variability and change in an ensemble of ocean reanalyses. *Clim. Dyn.* **49** 909–930.
- RISER, S. C., FREELAND, H. J., ROEMMICH, D., WIJFFELS, S., TROISI, A., BELBÉOCH, M., GILBERT, D., XU, J., POULIQUEN, S. et al. (2016). Fifteen years of ocean observations with the global Argo array. *Nat. Clim. Change* **6** 145–153. <https://doi.org/10.1038/nclimate2872>
- RISSE, M. D. (2016). Nonstationary spatial modeling, with emphasis on process convolution and covariate-driven approaches. ArXiv preprint. Available at [arXiv:1610.02447](https://arxiv.org/abs/1610.02447).
- RISSE, M. (2020). Nonstationary Bayesian modeling for a large data set of derived surface temperature return values. Available at [arXiv:2005.03658](https://arxiv.org/abs/2005.03658).
- RISSE, M. D. and TUREK, D. (2020). Bayesian inference for high-dimensional nonstationary Gaussian processes. *J. Stat. Comput. Simul.* **90** 2902–2928.
- ROBERTS, G. O. and ROSENTHAL, J. S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist.* **18** 349–367. [MR2749836 https://doi.org/10.1198/jcgs.2009.06134](https://doi.org/10.1198/jcgs.2009.06134)
- ROEMMICH, D. and GILSON, J. (2009). The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo program. *Progress in Oceanography* **82** 81–100.
- SAMPSON, P. D. and GUTTORP, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87** 108–119.
- STEIN, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer, New York. [MR1697409 https://doi.org/10.1007/978-1-4612-1494-6](https://doi.org/10.1007/978-1-4612-1494-6)
- TRENBERTH, K. E., FASULLO, J. T. and BALMASEDA, M. A. (2014). Earth's energy imbalance. *J. Climate* **27** 3129–3144.

- TUREK, D. and RISSE, M. D. (2019). BayesNSGP: Bayesian analysis of non-stationary Gaussian process models. Available at <https://CRAN.R-project.org/package=BayesNSGP>.
- VECCHIA, A. V. (1988). Estimation and model identification for continuous spatial processes. *J. Roy. Statist. Soc. Ser. B* **50** 297–312. [MR0964183](#)
- VON SCHUCKMANN, K. and LE TRAON, P.-Y. (2011). Schuckmann and Pierre-Yves Le Traon. How well can we derive global ocean indicators from Argo data? *Ocean Science* **7** 783–791.

BAYESIAN INFERENCE FOR BRAIN ACTIVITY FROM FUNCTIONAL MAGNETIC RESONANCE IMAGING COLLECTED AT TWO SPATIAL RESOLUTIONS

BY ANDREW S. WHITEMAN^{1,a}, ANDREAS J. BARTSCH^{2,d}, JIAN KANG^{1,b} AND
TIMOTHY D. JOHNSON^{1,c}

¹*Department of Biostatistics, University of Michigan School of Public Health, ^aawhitem@umich.edu, ^bjiankang@umich.edu, ^ctdjdj@umich.edu*

²*Radiologie Bamberg and Department of Neuroradiology, University of Heidelberg, ^ddr.bartsch@radiologie-bamberg.de*

Neuroradiologists and neurosurgeons increasingly opt to use functional magnetic resonance imaging (fMRI) to map functionally relevant brain regions for noninvasive presurgical planning and intraoperative neuronavigation. This application requires a high degree of spatial accuracy, but the fMRI signal-to-noise ratio (SNR) decreases as spatial resolution increases. In practice, fMRI scans can be collected at multiple spatial resolutions, and it is of interest to make more accurate inference on brain activity by combining data with different resolutions. To this end, we develop a new Bayesian model to leverage both better anatomical precision in high resolution fMRI and higher SNR in standard resolution fMRI. We assign a Gaussian process prior to the mean intensity function and develop an efficient, scalable posterior computation algorithm to integrate both sources of data. We draw posterior samples using an algorithm analogous to Riemann manifold Hamiltonian Monte Carlo in an expanded parameter space. We illustrate our method in analysis of presurgical fMRI data and show in simulation that it infers the mean intensity more accurately than alternatives that use either the high or standard resolution fMRI data alone.

REFERENCES

- ARCHIP, N., CLATZ, O., WHALEN, S., KACHER, D., FEDOROV, A., KOT, A., CHRISOCHOIDES, N., JOLESZ, F., GOLBY, A. et al. (2007). Non-rigid alignment of pre-operative MRI, fMRI, and DT-MRI with intra-operative MRI for enhanced visualization and navigation in image-guided neurosurgery. *NeuroImage* **35** 609–624.
- BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 825–848. [MR2523906 https://doi.org/10.1111/j.1467-9868.2008.00663.x](https://doi.org/10.1111/j.1467-9868.2008.00663.x)
- BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2012). Space-time data fusion under error in computer model output: An application to modeling air quality. *Biometrics* **68** 837–848. [MR3055188 https://doi.org/10.1111/j.1541-0420.2011.01725.x](https://doi.org/10.1111/j.1541-0420.2011.01725.x)
- BERTALANFFY, H., BENES, L., MIYAZAWA, T., ALBERTI, O., SIEGEL, A. M. and SURE, U. (2002). Cerebral cavernomas in the adult. Review of the literature and analysis of 72 surgically treated patients. *Neurosurg. Rev.* **25** 1–53.
- BODURKA, J., YE, F., PETRIDOU, N., MURPHY, K. and BANDETTINI, P. A. (2007). Mapping the MRI voxel volume in which thermal noise matches physiological noise—Implications for fMRI. *NeuroImage* **34** 542–549.
- BOWMAN, F. D., CAFFO, B., BASSETT, S. S. and KILTS, C. (2008). A Bayesian hierarchical framework for spatial modeling of fMRI data. *NeuroImage* **39** 146–156.
- BOWRING, A., TELSCHOW, F. J. E., SCHWARTZMAN, A. and NICHOLS, T. E. (2021). Confidence sets for Cohen’s d effect size images. *NeuroImage* **226** 117477. <https://doi.org/10.1016/j.neuroimage.2020.117477>
- CORDELLA, R., ACERBI, F., BROGGI, M., VILATI, D., NAZZI, V., SCHIARITI, M., TRINGALI, G., FERROLI, P., FRANZINI, A. et al. (2013). Intraoperative neurophysiological monitoring of the cortico-spinal tract in image-guided mini-invasive neurosurgery. *Clin. Neurophysiol.* **124** 1244–1254.

Key words and phrases. Imaging statistics, Gaussian process, Bayesian nonparametrics, data integration, presurgical fMRI.

- COX, R., ASHBURNER, J., BREMAN, H., FISSELL, K., HASELGROVE, C., HOLMES, C., LANCASTER, J., REX, D., SMITH, S. et al. (2004). A (sort of) new image data format standard: NIFTI-1: WE 150. *NeuroImage* **22**.
- CRESSIE, N. and GLONEK, G. (1984). Median based covariogram estimators reduce bias. *Statist. Probab. Lett.* **2** 299–304. [MR0777844 https://doi.org/10.1016/0167-7152\(84\)90069-5](https://doi.org/10.1016/0167-7152(84)90069-5)
- DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. [MR3538706 https://doi.org/10.1080/01621459.2015.1044091](https://doi.org/10.1080/01621459.2015.1044091)
- DIGGLE, P. J. (1981). Binary mosaics and the spatial pattern of heather. *Biometrics* **37** 531–539.
- DURNEZ, J., MOERKERKE, B., BARTSCH, A. and NICHOLS, T. E. (2013). Alternative-based thresholding with application to presurgical fMRI. *Cogn. Affect. Behav. Neurosci.* **13** 703–713.
- FINLEY, A. O., DATTA, A., COOK, B. D., MORTON, D. C., ANDERSEN, H. E. and BANERJEE, S. (2019). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *J. Comput. Graph. Statist.* **28** 401–414. [MR3974889 https://doi.org/10.1080/10618600.2018.1537924](https://doi.org/10.1080/10618600.2018.1537924)
- FRIGO, M. and JOHNSON, S. G. (2005). The design and implementation of FFTW3. *Proc. IEEE* **93** 216–231. Special issue on “Program Generation, Optimization, and Platform Adaptation”.
- FUENTES, M. and RAFTERY, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61** 36–45. [MR2129199 https://doi.org/10.1111/j.0006-341X.2005.030821.x](https://doi.org/10.1111/j.0006-341X.2005.030821.x)
- GELFAND, A. E., ZHU, L. and CARLIN, B. P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics* **2** 31–45.
- GELMAN, A., RUBIN, D. B. et al. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
- GENOVESE, C. R., LAZAR, N. A. and NICHOLS, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* **15** 870–878.
- GIROLAMI, M. and CALDERHEAD, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 123–214. [MR2814492 https://doi.org/10.1111/j.1467-9868.2010.00765.x](https://doi.org/10.1111/j.1467-9868.2010.00765.x)
- GROVES, A. R., CHAPPELL, M. A. and WOOLRICH, M. W. (2009). Combined spatial and non-spatial prior for inference on MRI time-series. *NeuroImage* **45** 795–809. <https://doi.org/10.1016/j.neuroimage.2008.12.027>
- GUENNEBAUD, G., JACOB, B. et al. (2010). Eigen v3. Available at <http://eigen.tuxfamily.org>.
- HALLER, S. and BARTSCH, A. J. (2009). Pitfalls in fMRI. *Eur. Radiol.* **19** 2689–2706.
- JENKINSON, M., BECKMANN, C. F., BEHRENS, T. E., WOOLRICH, M. W. and SMITH, S. M. (2012). Fsl. *NeuroImage* **62** 782–790.
- JOHNSON, S. G. The NLOpt nonlinear optimization package. Available at <http://github.com/stevengj/nlopt>.
- JOVČEVSKA, I., KOČEVAR, N. and KOMEL, R. (2013). Glioma and glioblastoma-how much do we (not) know? *Mol. Clin. Oncol.* **1** 935–941.
- LARGE, I., BRIDGE, H., AHMED, B., CLARE, S., KOLASINSKI, J., LAM, W., MILLER, K., DYRBY, T. B., PARKER, A. et al. (2016). Individual differences in the alignment of structural and functional markers of the V5/MT complex in primates. *Cereb. Cortex* **26** 3928–3944.
- LIU, Z., BERROCAL, V. J., BARTSCH, A. J. and JOHNSON, T. D. (2016). Pre-surgical fMRI data analysis using a spatially adaptive conditionally autoregressive model. *Bayesian Anal.* **11** 599–625. [MR3472004 https://doi.org/10.1214/15-BA972](https://doi.org/10.1214/15-BA972)
- LIU, Z., BARTSCH, A. J., BERROCAL, V. J. and JOHNSON, T. D. (2019). A mixed-effects, spatially varying coefficients model with application to multi-resolution functional magnetic resonance imaging data. *Stat. Methods Med. Res.* **28** 1203–1215. [MR3934644 https://doi.org/10.1177/0962280217752378](https://doi.org/10.1177/0962280217752378)
- MARDIA, K. V. and MARSHALL, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71** 135–146. [MR0738334 https://doi.org/10.1093/biomet/71.1.135](https://doi.org/10.1093/biomet/71.1.135)
- MÜLLER, P., PARMIGIANI, G. and RICE, K. (2007). FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics 8. Oxford Sci. Publ.* 349–370. Oxford Univ. Press, Oxford. [MR2433200](https://doi.org/10.1017/0962280217752378)
- NIMSKY, C., GANSLANDT, O., BUCHFELDER, M. and FAHLBUSCH, R. (2006). Intraoperative visualization for resection of gliomas: The role of functional neuronavigation and intraoperative 1.5 T MRI. *Neurol. Res.* **28** 482–487.
- PENNY, W. D., TRUJILLO-BARRETO, N. J. and FRISTON, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage* **24** 350–362.
- POWELL, M. J. D. (1994). A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in Optimization and Numerical Analysis (Oaxaca, 1992)*. *Math. Appl.* **275** 51–67. Kluwer Academic, Dordrecht. [MR1282734](https://doi.org/10.1007/978-94-017-1282-7_3)
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR2514435](https://doi.org/10.1017/9780521876223)

- REUTER, M., ROSAS, H. D. and FISCHL, B. (2010). Highly accurate inverse consistent registration: A robust approach. *NeuroImage* **53** 1181–1196.
- RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields: Theory and Applications. Monographs on Statistics and Applied Probability* **104**. CRC Press/CRC, Boca Raton, FL. MR2130347 <https://doi.org/10.1201/9780203492024>
- SEEGER, M., WILLIAMS, C. and LAWRENCE, N. (2003). Fast forward selection to speed up sparse Gaussian process regression. Technical report.
- SIDÉN, P., EKLUND, A., BOLIN, D. and VILLANI, M. (2017). Fast Bayesian whole-brain fMRI analysis with spatial 3D priors. *NeuroImage* **146** 211–225. <https://doi.org/10.1016/j.neuroimage.2016.11.040>
- SILVA, M. A., SEE, A. P., ESSAYED, W. I., GOLBY, A. J. and TIE, Y. (2018). Challenges and techniques for presurgical brain mapping with functional MRI. *NeuroImage Clin.* **17** 794–803. <https://doi.org/10.1016/j.nicl.2017.12.008>
- STEIN, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging. Springer Series in Statistics*. Springer, New York. MR1697409 <https://doi.org/10.1007/978-1-4612-1494-6>
- STEIN, M. L. (2007). Spatial variation of total column ozone on a global scale. *Ann. Appl. Stat.* **1** 191–210. MR2393847 <https://doi.org/10.1214/07-AOAS106>
- STIPPICH, C. (2015). *Clinical Functional MRI: Presurgical Functional Neuroimaging*. Springer, Berlin.
- VAN HORN, J. D. and TOGA, A. W. (2009). Multi-site neuroimaging trials. *Curr. Opin. Neurol.* **22** 370.
- WHITEMAN, A. S., BARTSCH, A. J. KANG, J. and JOHNSON, T. D. (2022). Supplement to “Bayesian inference for brain activity from functional magnetic resonance imaging collected at two spatial resolutions.” <https://doi.org/10.1214/22-AOAS1606SUPPA>
- WHITEMAN, A. S., BARTSCH, A. J. KANG, J. and JOHNSON, T. D. (2022). Source code companion to “Bayesian inference for brain activity from functional magnetic resonance imaging collected at two spatial resolutions.” <https://doi.org/10.1214/22-AOAS1606SUPPB>
- WOOD, A. T. A. and CHAN, G. (1994). Simulation of stationary Gaussian processes in $[0, 1]^d$. *J. Comput. Graph. Statist.* **3** 409–432. MR1323050 <https://doi.org/10.2307/1390903>
- WOOLRICH, M. W., RIPLEY, B. D., BRADY, M. and SMITH, S. M. (2001). Temporal autocorrelation in univariate linear modeling of fMRI data. *NeuroImage* **14** 1370–1386.
- WOOLRICH, M. W., JENKINSON, M., BRADY, J. M. and SMITH, S. M. (2004). Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE Trans. Med. Imag.* **23** 213–231.
- YUE, Y., LOH, J. M. and LINDQUIST, M. A. (2010). Adaptive spatial smoothing of fMRI images. *Stat. Interface* **3** 3–13. MR2609707 <https://doi.org/10.4310/SII.2010.v3.n1.a1>
- ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261. MR2054303 <https://doi.org/10.1198/016214504000000241>

TESTING FOR DIFFERENTIAL ABUNDANCE IN COMPOSITIONAL COUNTS DATA, WITH APPLICATION TO MICROBIOME STUDIES

BY BARAK BRILL^{1,a}, AMNON AMIR^{2,c} AND RUTH HELLER^{1,b}

¹Department of Statistics and Operations Research, Tel Aviv University, ^abarakbri@mail.tau.ac.il, ^bruheller@gmail.com

²Microbiome center, The Chaim Sheba Medical Center, ^camnonim@gmail.com

Identifying which taxa in our microbiota are associated with traits of interest is important for advancing science and health. However, the identification is challenging because the measured vector of taxa counts (by amplicon sequencing) is compositional, so a change in the abundance of one taxon in the microbiota induces a change in the number of sequenced counts across all taxa. The data are typically sparse, with many zero counts present either due to biological variance or limited sequencing depth. We examine the case of Crohn's disease, where the microbial load changes substantially with the disease. For this representative example of a highly compositional setting, we show existing methods designed to identify differentially abundant taxa may have an inflated number of false positives. We introduce a novel nonparametric approach that provides valid inference, even when the fraction of zero counts is substantial. Our approach uses a set of reference taxa that are non-differentially abundant which can be estimated from the data or from outside information. Our approach also allows for a novel type of testing: multivariate tests of differential abundance over a focused subset of the taxa. Genera-level multivariate testing discovers additional genera as differentially abundant by avoiding agglomeration of taxa.

REFERENCES

- AITCHISON, J. (1982). The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B* **44** 139–177. [MR0676206](#)
- AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability*. CRC Press, London. [MR0865647](#) <https://doi.org/10.1007/978-94-009-4109-0>
- AMIR, A., McDONALD, D., NAVAS-MOLINA, J. A., KOPYLOVA, E., MORTON, J. T., XU, Z. Z., KIGHTLEY, E. P., THOMPSON, L. R., HYDE, E. R. et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* **2** e00191-16.
- ANDERSON, M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26** 32–46.
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. [MR1325392](#)
- BRILL, B., AMIR, A. and HELLER, R. (2022). Supplement to “Testing for differential abundance in compositional counts data, with application to microbiome studies.” <https://doi.org/10.1214/22-AOAS1607SUPPA>, <https://doi.org/10.1214/22-AOAS1607SUPPB>, <https://doi.org/10.1214/22-AOAS1607SUPPC>, <https://doi.org/10.1214/22-AOAS1607SUPPD>, <https://doi.org/10.1214/22-AOAS1607SUPPE>
- CALGARO, M., ROMUALDI, C., WALDRON, L., RISSO, D. and VITULO, N. (2020). Assessment of single cell rna-seq statistical methods on microbiome data. *BioRxiv*.
- CALLAHAN, B. J., MCMURDIE, P. J., ROSEN, M. J., HAN, A. W., JOHNSON, A. J. A. and HOLMES, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13** 581–583.
- DESANTIS, T. Z., HUGENHOLTZ, P., LARSEN, N., ROJAS, M., BRODIE, E. L., KELLER, K., HUBER, T., DALEVI, D., HU, P. et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72** 5069–5072.
- FERNANDES, A. D., MACKLAIM, J. M., LINN, T. G., REID, G. and GLOOR, G. B. (2013). ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-seq. *PLoS ONE* **8** e67019.

Key words and phrases. Compositional bias, analysis of composition, normalization, rarefaction, nonparametric tests.

- FORBES, C., EVANS, M., HASTINGS, N. and PEACOCK, B. (2011). *Statistical Distributions*. Wiley, Hoboken, NJ. [MR2964192](#)
- GEVERS, D., KNIGHT, R., PETROSINO, J. F., HUANG, K., MCGUIRE, A. L., BIRREN, B. W., NELSON, K. E., WHITE, O., METHE, B. A. et al. (2012). The human microbiome project: A community resource for the healthy human microbiome. *PLoS Biol.* **10** e1001377.
- GLOOR, G. B., MACKLAIM, J. M., PAWLOWSKY-GLAHN, V. and EGOZCUE, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8** 2224.
- GOODRICH, J. K., DI RIENZI, S. C., POOLE, A. C., KOREN, O., WALTERS, W. A., CAPORASO, J. G., KNIGHT, R. and LEY, R. E. (2014). Conducting a microbiome study. *Cell* **158** 250–262.
- GRETTON, A., FUKUMIZU, K., TEO, C., SONG, L., SCHÖLKOPF, B. and SMOLA, A. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems* 20, Red Hook, NY, USA 585–592. Max-Planck-Gesellschaft, Curran.
- GUO, X., ZHANG, X., QIN, Y., LIU, Y.-X., ZHANG, J., ZHANG, N., WU, K., QU, B., HE, Z. et al. (2020). Host-associated quantitative abundance profiling reveals the microbial load variation of root microbiome. *Plant Commun.* **1** 100003.
- HAMADY, M. and KNIGHT, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res.* **19** 1141–1152.
- HAWINKEL, S., MATTIELLO, F., BIJNENS, L. and THAS, O. (2019). A broken promise: Microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.* **20** 210–221.
- HELLER, R., HELLER, Y. and GORFINE, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100** 503–510. [MR3068450](#) <https://doi.org/10.1093/biomet/ass070>
- HOMMEL, G. and KROPP, S. (2005). Tests for differentiation in gene expression using a data-driven order or weights for hypotheses. *Biom. J.* **47** 554–562. [MR2190478](#) <https://doi.org/10.1002/bimj.200410118>
- JIAN, C., LUUKKONEN, P., YKI-JÄRVINEN, H., SALONEN, A. and KORPELA, K. (2020). Quantitative pcr provides a simple and accessible method for quantitative microbiota profiling. *PLoS ONE* **15** e0227285.
- JIANG, L., AMIR, A., MORTON, J. T., HELLER, R., ARIAS-CASTRO, E. and KNIGHT, R. (2017). Discrete false-discovery rate improves identification of differentially abundant microbes. *mSystems* **2** e00092-17.
- JIANG, S.-Q., YU, Y.-N., GAO, R.-W., WANG, H., ZHANG, J., LI, R., LONG, X.-H., SHEN, Q.-R., CHEN, W. et al. (2019). High-throughput absolute quantification sequencing reveals the effect of different fertilizer applications on bacterial community in a tomato cultivated coastal saline soil. *Sci. Total Environ.* **687** 601–609.
- KALIVODOVA, A., HRON, K., FILZMOSER, P., NAJDEKR, L., JANECKOVA, H. and ADAM, T. (2015). PLS-DA for compositional data with application to metabolomics. *J. Chemom.* **29** 21–28.
- KAUL, A., MANDAL, S., DAVIDOV, O. and PEDDADA, S. D. (2017). Analysis of microbiome data in the presence of excess zeros. *Front. Microbiol.* **8** 2114.
- KONG, J., LIU, X., WANG, L., HUANG, H., OU, D., GUO, J., LAWS, E. A. and HUANG, B. (2021). Patterns of relative and quantitative abundances of marine bacteria in surface waters of the subtropical northwest Pacific Ocean estimated with high-throughput quantification sequencing. *Front. Microbiol.* **11** 599614.
- KUMAR, M. S., SLUD, E. V., OKRAH, K., HICKS, S. C., HANNENHALLI, S. and BRAVO, H. C. (2018). Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics* **19** 799.
- LOVE, M. I., HUBER, W. and ANDERS, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15** 550.
- LUZ, C. M. (2019). Statistical analysis of metagenomics data. *Genomics Inform.* **17** e6.
- MANDAL, S., TREUREN, W. V., WHITE, R. A., EGGESBØ, M., KNIGHT, R. and PEDDADA, S. D. (2015). Analysis of composition of microbiomes: A novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26** 27663.
- MCDONALD, D., HYDE, E., DEBELIUS, J. W., MORTON, J. T., GONZALEZ, A., ACKERMANN, G., AKSENOV, A. A., BEHSAZ, B., BRENNAN, C. et al. (2018). American gut: An open platform for citizen science microbiome research. *mSystems* **3** e00031-18.
- MCMURDIE, P. J. and HOLMES, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10** e1003531.
- MORTON, J. T., MAROTZ, C., WASHBURNE, A., SILVERMAN, J., ZARAMELA, L. S., EDLUND, A., ZENGLER, K. and KNIGHT, R. (2019). Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* **10** 2719.
- NELSON, M. C., MORRISON, H. G., BENJAMINO, J., GRIM, S. L. and GRAF, J. (2014). Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS ONE* **9** e94249.
- PAULSON, J. N., POP, M. and BRAVO, H. C. (2013). metagenomeSeq: Statistical analysis for sparse high-throughput sequencing. Bioconductor package.
- PAULSON, J. N., STINE, O. C., BRAVO, H. C. and POP, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10** 1200–1202.

- QUINN, T. P., ERB, I., GLOOR, G., NOTREDAME, C., RICHARDSON, M. F. and CROWLEY, T. M. (2019). A field guide for the compositional analysis of any-omics data. *GigaScience* **8** giz107.
- RISSO, D., PERRAUDEAU, F., GRIBKOVA, S., DUDOIT, S. and VERT, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9** 284.
- 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>
- SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754. MR0897872 <https://doi.org/10.1093/biomet/73.3.751>
- STAEMLER, F., GLAESNER, J., HIERGEIST, A., HOLLER, E., WEBER, D., OEFNER, P. J., GESSNER, A. and SPANG, R. (2016). Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* **4** 28.
- SUNAGAWA, S., COELHO, L. P., CHAFFRON, S., KULTIMA, J. R., LABADIE, K., SALAZAR, G., DJAHAN-SCHIRI, B., ZELLER, G., MENDE, D. R. et al. (2015). Structure and function of the global ocean microbiome. *Science* **348** 1261359.
- SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265. MR2752127 <https://doi.org/10.1214/09-AOAS312>
- TKACZ, A., HORTALA, M. and POOLE, P. S. (2018). Absolute quantitation of microbiota abundance in environmental samples. *Microbiome* **6** 110.
- TORGERSON, W. S. (1952). Multidimensional scaling. I. Theory and method. *Psychometrika* **17** 401–419. MR0054219 <https://doi.org/10.1007/BF02288916>
- TSAGRIS, M., ALENAZI, A., VERRON, K.-M. and PANDIS, N. (2020). Hypothesis testing for two population means: Parametric or non-parametric test? *J. Stat. Comput. Simul.* **90** 252–270. MR4038806 <https://doi.org/10.1080/00949655.2019.1677659>
- VANDEPUTTE, D., KATHAGEN, G., D’HOE, K., VIEIRA-SILVA, S., VALLES-COLOMER, M., SABINO, J., WANG, J., TITO, R. Y., DE COMMER, L. et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551** 507–511.
- VAN DEN BERGE, K., PERRAUDEAU, F., SONESON, C., LOVE, M. I., RISSO, D., VERT, J.-P., ROBINSON, M. D., DUDOIT, S. and CLEMENT, L. (2018). Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* **19** 24.
- VIEIRA, M. C. (2020). Evolution of adaptability and the immune response to influenza and HIV. Ph. D. thesis, The Univ. Chicago.
- VIEIRA-SILVA, S., SABINO, J., VALLES-COLOMER, M., FALONY, G., KATHAGEN, G., CAENEPEEL, C., CLEYNEN, I., VAN DER MERWE, S., VERMEIRE, S. et al. (2019). Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. *Nat. Microbiol.* **4** 1826–1831.
- WANG, Q., GARRITY, G. M., TIEDJE, J. M. and COLE, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73** 5261–5267.
- WEISS, S., XU, Z. Z., PEDDADA, S., AMIR, A., BITTINGER, K., GONZALEZ, A., LOZUPONE, C., ZAN-EVELD, J. R., VAZQUEZ-BAEZA, Y. et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5** 27.
- WU, J. R., MACKLAIM, J. M., GENGE, B. L. and GLOOR, G. B. (2017). Finding the centre: Corrections for asymmetry in high-throughput sequencing datasets. Available at [arXiv:1704.01841](https://arxiv.org/abs/1704.01841).
- XU, L., PATERSON, A. D., TURPIN, W. and XU, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PLoS ONE* **10** e0129606.

MAPPING INTERSTELLAR DUST WITH GAUSSIAN PROCESSES

BY ANDREW C. MILLER^{1,a}, LAUREN ANDERSON^{4,e}, BORIS LEISTEDT^{4,f}, JOHN P. CUNNINGHAM^{2,c}, DAVID W. HOGG^{3,d} AND DAVID M. BLEI^{1,b}

¹Data Science Institute, Columbia University, ^aam5171@columbia.edu, ^bdavid.blei@columbia.edu

²Department of Statistics, Columbia University, ^cjpc2181@columbia.edu

³Department of Physics, New York University, ^ddavid.hogg@nyu.edu

⁴Center for Computational Astrophysics, Flatiron Institute, ^eanders.astro@gmail.com, ^fboris.leistedt@gmail.com

Interstellar dust corrupts nearly every stellar observation and accounting for it is crucial to measuring physical properties of stars. We model the dust distribution as a spatially varying latent field with a Gaussian process (GP) and develop a likelihood model and inference method that scales to millions of astronomical observations. Modeling interstellar dust is complicated by two factors. The first is *integrated observations*. The data come from a vantage point on Earth, and each observation is an integral of the unobserved function along our line of sight, resulting in a complex likelihood and a more difficult inference problem than in classical GP inference. The second complication is *scale*; stellar catalogs have millions of observations. To address these challenges, we develop ZIGGY, a scalable approach to GP inference with integrated observations based on stochastic variational inference. We study ZIGGY on synthetic data and the Ananke dataset, a high-fidelity mechanistic model of the Milky Way with millions of stars. ZIGGY reliably infers the spatial dust map with well-calibrated posterior uncertainties.

REFERENCES

- AGHANIM, N. et al. (2018). Planck 2018 results. VI. Cosmological parameters. E-prints. Available at [arXiv:1807.06209](https://arxiv.org/abs/1807.06209).
- AGUADO, D. S., AHUMADA, R., ALMEIDA, A., ANDERSON, S. F., ANDREWS, B. H., ANGUIANO, B., ORTÍZ, E. A., ARAGÓN-SALAMANCA, A., ARGUDO-FERNÁNDEZ, M. et al. (2019). The fifteenth data release of the Sloan Digital Sky Surveys: First release of MaNGA-derived quantities, data visualization tools, and Stellar Library. *Astrophys. J., Suppl. Ser.* **240** 23.
- AMARI, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Comput.* **10** 251–276.
- AMBIKASARAN, S., FOREMAN-MACKEY, D., GREENGARD, L., HOGG, D. W. and O’NEIL, M. (2014). Fast direct methods for Gaussian processes. Preprint. Available at [arXiv:1403.6015](https://arxiv.org/abs/1403.6015).
- BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. *Monographs on Statistics and Applied Probability* **135**. CRC Press, Boca Raton, FL. MR3362184
- BERNARDO, J., BAYARRI, M., BERGER, J., DAWID, A., HECKERMAN, D., SMITH, A. and WEST, M. (2003). Non-centered parameterisations for hierarchical models and data augmentation. In *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting* **307**. Oxford Univ. Press, USA.
- 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>
- BROWN, A., VALLENARI, A., PRUSTI, T., DE BRUIJNE, J., BABUSIAUX, C., BAILER-JONES, C., BIERMANN, M., EVANS, D. W., EYER, L. et al. (2018). Gaia data release 2-summary of the contents and survey properties. *Astron. Astrophys.* **616** A1.
- BURT, D. R., RASMUSSEN, C. E. and VAN DER WILK, M. (2019). Rates of convergence for sparse variational Gaussian process regression. Preprint. Available at [arXiv:1903.03571](https://arxiv.org/abs/1903.03571).
- CHEN, B. Q., HUANG, Y., HOU, L. G., TIAN, H., LI, G. X., YUAN, H. B., WANG, H. F., WANG, C., TIAN, Z. J. et al. (2019). The galactic spiral structure as revealed by O- and early B-type stars. *Mon. Not. R. Astron. Soc.* **487** 1400–1409. <https://doi.org/10.1093/mnras/stz1357>

- CRESSIE, N. (1990). The origins of Kriging. *Math. Geol.* **22** 239–252. [MR1047810](#) <https://doi.org/10.1007/BF00889887>
- CRESSIE, N. (1992). Statistics for spatial data. *Terra Nova* **4** 613–617.
- DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. [MR3538706](#) <https://doi.org/10.1080/01621459.2015.1044091>
- EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. [MR2431866](#) <https://doi.org/10.1214/07-STS236>
- EFRON, B. (2019). Bayes, oracle Bayes and empirical Bayes. *Statist. Sci.* **34** 177–201. [MR3983318](#) <https://doi.org/10.1214/18-STS674>
- FLEWELLING, H. (2016). The pan-STARRS 1 medium deep field variable star catalog. In *American Astronomical Society Meeting Abstracts #227. American Astronomical Society Meeting Abstracts* **227** 144.25.
- FOREMAN-MACKEY, D., AGOL, E., AMBIKASARAN, S. and ANGUS, R. (2017). Fast and scalable Gaussian process modeling with applications to astronomical time series. *Astron. J.* **154** 220.
- GENTON, M. G. (2002). Classes of kernels for machine learning: A statistics perspective. *J. Mach. Learn. Res.* **2** 293–312. [MR1904760](#) <https://doi.org/10.1162/15324430260185637>
- GEORGELIN, Y. M. and GEORGELIN, Y. P. (1976). The spiral structure of our Galaxy determined from H II regions. *Astron. Astrophys.* **49** 57–79.
- GNEITING, T. (2002). Compactly supported correlation functions. *J. Multivariate Anal.* **83** 493–508. [MR1945966](#) <https://doi.org/10.1006/jmva.2001.2056>
- GREEN, G. M., SCHLAFLY, E. F., FINKBEINER, D., RIX, H.-W., MARTIN, N., BURGETT, W., DRAPER, P. W., FLEWELLING, H., HODAPP, K. et al. (2018). Galactic reddening in 3D from stellar photometry—an improved map. *Mon. Not. R. Astron. Soc.* **478** 651–666.
- GREEN, G. M., SCHLAFLY, E. F., ZUCKER, C., SPEAGLE, J. S. and FINKBEINER, D. P. (2019). A 3D dust map based on gaia, pan-STARRS 1 and 2MASS. Preprint. Available at [arXiv:1905.02734](#).
- HEATON, M. J., DATTA, A., FINLEY, A. O. et al. (2019). A case study competition among methods for analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* **24** 398–425. [MR3996451](#) <https://doi.org/10.1007/s13253-018-00348-w>
- HENSMAN, J., FUSI, N. and LAWRENCE, N. D. (2013). Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence* 282–290. AUAI Press.
- HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.* **14** 1303–1347. [MR3081926](#)
- HOPKINS, P. F., WETZEL, A., KEREŠ, D., FAUCHER-GIGUÈRE, C.-A., QUATAERT, E., BOYLAN-KOLCHIN, M., MURRAY, N., HAYWARD, C. C., GARRISON-KIMMEL, S. et al. (2018). FIRE-2 simulations: Physics versus numerics in galaxy formation. *Mon. Not. R. Astron. Soc.* **480** 800–863. <https://doi.org/10.1093/mnras/sty1690>
- 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.
- KLYPIN, A., KRAVTSOV, A. V., VALENZUELA, O. and PRADA, F. (1999). Where are the missing galactic satellites? *Astrophys. J.* **522** 82–92. <https://doi.org/10.1086/307643>
- KRIGE, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *J. S. Afr. Inst. Min. Metall.* **52** 119–139.
- LE, Q., SARLÓS, T. and SMOLA, A. (2013). Fastfood-computing Hilbert space expansions in loglinear time. In *International Conference on Machine Learning* 244–252.
- LEIKE, R. and ENSSLIN, T. (2019). Charting nearby dust clouds using Gaia data only. Preprint. Available at [arXiv:1901.05971](#).
- MATÉRN, B. (1960). Spatial variation: Stochastic models and their applications to some problems in forest surveys and other sampling investigations. *Meddelanden från Statens Skogsforskningsinstitut* **49** 1–144.
- MATHERON, G. (1963). Principles of geostatistics. *Econ. Geol.* **58** 1246–1266.
- MATHERON, G. (1973). The intrinsic random functions and their applications. *Adv. in Appl. Probab.* **5** 439–468. [MR0356209](#) <https://doi.org/10.2307/1425829>
- MATHIS, J. S. (1990). Interstellar dust and extinction. *Annu. Rev. Astron. Astrophys.* **28** 37–70.
- MILLER, A. C., ANDERSON, L., LEISTEDT, B., CUNNINGHAM, J. P., HOGG, D. W. and BLEI, D. M. (2022). Supplement to “Mapping interstellar dust with Gaussian processes.” <https://doi.org/10.1214/22-AOAS1608SUPPA>, <https://doi.org/10.1214/22-AOAS1608SUPPB>
- MURRAY, I. and ADAMS, R. P. (2010). Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems* 1732–1740.
- QUIÑONERO-CANDELA, J. and RASMUSSEN, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.* **6** 1939–1959. [MR2249877](#)

- RAHIMI, A. and RECHT, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems* 1177–1184.
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR2514435](#)
- REZAEI KH., S., BAILER-JONES, C., HANSON, R. and FOUESNEAU, M. (2017). Inferring the three-dimensional distribution of dust in the Galaxy with a non-parametric method-preparing for Gaia. *Astron. Astrophys.* **598** A125.
- REZAEI KH., S., BAILER-JONES, C. A. L., HOGG, D. W. and SCHULTHEIS, M. (2018). Detection of the Milky Way spiral arms in dust from 3D mapping. *Astron. Astrophys.* **618** A168. <https://doi.org/10.1051/0004-6361/201833284>
- ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. [MR0042668](#) <https://doi.org/10.1214/aoms/1177729586>
- 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>
- SANDERSON, R. E., WETZEL, A., LOEBMAN, S., SHARMA, S., HOPKINS, P. F., GARRISON-KIMMEL, S., FAUCHER-GIGUÈRE, C.-A., KEREŠ, D. and QUATAERT, E. (2018). Synthetic Gaia surveys from the FIRE cosmological simulations of Milky-Way-mass galaxies. E-prints. Available at [arXiv:1806.10564](#).
- SCHLEGEL, D. J., FINKBEINER, D. P. and DAVIS, M. (1998). Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds. *Astrophys. J.* **500** 525.
- SNELSON, E. and GHARAMANI, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems* 1257–1264.
- TITSIAS, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics* 567–574.
- WAINWRIGHT, M. J., JORDAN, M. I. et al. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.
- WETZEL, A. R., HOPKINS, P. F., HOON KIM, J., FAUCHER-GIGUÈRE, C.-A., KEREŠ, D. and QUATAERT, E. (2016). Reconciling Dwarf Galaxies with Λ CMD cosmology: Simulating a realistic population of Satellites Around a Milky Way Mass Galaxy. *Astrophys. J.* **827** L23. <https://doi.org/10.3847/2041-8205/827/2/L23>
- YANG, T., LI, Y.-F., MAHDAVI, M., JIN, R. and ZHOU, Z.-H. (2012). Nyström method vs random Fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems* 476–484.

MODELLING EXTREMES OF SPATIAL AGGREGATES OF PRECIPITATION USING CONDITIONAL METHODS

BY JORDAN RICHARDS^{1,a}, JONATHAN A. TAWN^{1,b} AND SIMON BROWN^{2,c}

¹STOR-i Centre for Doctoral Training, Department of Mathematics and Statistics, Lancaster University,
jordan.richards@kaust.edu.sa, j.tawn@lancaster.ac.uk

²Hadley Centre for Climate Prediction and Research, Met Office, simon.brown@metoffice.gov.uk

Inference on the extremal behaviour of spatial aggregates of precipitation is important for quantifying river flood risk. There are two classes of previous approach, with one failing to ensure self-consistency in inference across different regions of aggregation and the other imposing highly restrictive assumptions. To overcome these issues, we propose a model for high-resolution precipitation data from which we can simulate realistic fields and explore the behaviour of spatial aggregates. Recent developments have seen spatial extensions of the Heffernan and Tawn (*J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** (2004) 497–546) model for conditional multivariate extremes which can handle a wide range of dependence structures. Our contribution is twofold: extensions and improvements of this approach and its model inference for high-dimensional data and a novel framework for deriving aggregates addressing edge effects and subregions without rain. We apply our modelling approach to gridded East Anglia, UK precipitation data. Return-level curves for spatial aggregates over different regions of various sizes are estimated and shown to fit very well to the data.

REFERENCES

- BENOIT, L., ALLARD, D. and MARIETHOZ, G. (2018). Stochastic rainfall modeling at sub-kilometer scale. *Water Resour. Res.* **54** 4108–4130.
- BLANCHET, J. and DAVISON, A. C. (2011). Spatial modeling of extreme snow depth. *Ann. Appl. Stat.* **5** 1699–1725. [MR2884920](#) <https://doi.org/10.1214/11-AOAS464>
- BREIMAN, L. (1965). On some limit theorems similar to the arc-sin law. *Theory Probab. Appl.* **10** 323–331. [MR0184274](#)
- BROWN, S. J. (2020). Future changes in heatwave severity, duration and frequency due to climate change for the most populous cities. *Weather Clim. Extrem.* **30** 100278. <https://doi.org/10.1016/j.wace.2020.100278>
- BUISHAND, T. A., DE HAAN, L. and ZHOU, C. (2008). On spatial extremes: With application to a rainfall problem. *Ann. Appl. Stat.* **2** 624–642. [MR2524349](#) <https://doi.org/10.1214/08-AOAS159>
- CASTRO-CAMILO, D. and HUSER, R. (2020). Local likelihood estimation of complex tail dependence structures, applied to U.S. precipitation extremes. *J. Amer. Statist. Assoc.* **115** 1037–1054. [MR4143448](#) <https://doi.org/10.1080/01621459.2019.1647842>
- COLES, S. G. (1993). Regional modelling of extreme storms via max-stable processes. *J. Roy. Statist. Soc. Ser. B* **55** 797–816. [MR1229882](#)
- COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer, London. [MR1932132](#) <https://doi.org/10.1007/978-1-4471-3675-0>
- COLES, S. G. and TAWN, J. A. (1996). Modelling extremes of the areal rainfall process. *J. Roy. Statist. Soc. Ser. B* **58** 329–347. [MR1377836](#)
- 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>
- DAVISON, A. C. and SMITH, R. L. (1990). Models for exceedances over high thresholds (with discussion). *J. Roy. Statist. Soc. Ser. B* **52** 393–442. [MR1086795](#)
- DE FONDEVILLE, R. and DAVISON, A. C. (2022). Functional peaks-over-threshold analysis. *J. Roy. Statist. Soc. Ser. B*. Available at <https://doi.org/10.1111/rssb.12498>

- ENGELKE, S., DE FONDEVILLE, R. and OESTING, M. (2019). Extremal behaviour of aggregated data with an application to downscaling. *Biometrika* **106** 127–144. [MR3912387](#) <https://doi.org/10.1093/biomet/asy052>
- FASIOLO, M., WOOD, S. N., ZAFFRAN, M., NEDELLEC, R. and GOUDE, Y. (2021). Fast calibrated additive quantile regression. *J. Amer. Statist. Assoc.* **116** 1402–1412. [MR4309281](#) <https://doi.org/10.1080/01621459.2020.1725521>
- FERREIRA, A., DE HAAN, L. and ZHOU, C. (2012). Exceedance probability of the integral of a stochastic process. *J. Multivariate Anal.* **105** 241–257. [MR2877515](#) <https://doi.org/10.1016/j.jmva.2011.08.020>
- HARRISON, D. L., DRISCOLL, S. J. and KITCHEN, M. (2000). Improving precipitation estimates from weather radar using quality control and correction techniques. *Meteorol. Appl.* **7** 135–144.
- HEFFERNAN, J. E. and TAWN, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 497–546. [MR2088289](#) <https://doi.org/10.1111/j.1467-9868.2004.02050.x>
- HUSER, R. and DAVISON, A. C. (2013). Composite likelihood estimation for the Brown–Resnick process. *Biometrika* **100** 511–518. [MR3068451](#) <https://doi.org/10.1093/biomet/ass089>
- HUSER, R. and GENTON, M. G. (2016). Non-stationary dependence structures for spatial extremes. *J. Agric. Biol. Environ. Stat.* **21** 470–491. [MR3542082](#) <https://doi.org/10.1007/s13253-016-0247-4>
- HUSER, R. and WADSWORTH, J. L. (2022). Advances in statistical modeling of spatial extremes. *Wiley Interdiscip. Rev.: Comput. Stat.* **14** Paper No. e1537, 28. [MR4376783](#) <https://doi.org/10.1002/wics.1537>
- JOE, H. (1997). *Multivariate Models and Dependence Concepts. Monographs on Statistics and Applied Probability* **73**. CRC Press, London. [MR1462613](#) <https://doi.org/10.1201/b13150>
- KENDON, E. J., ROBERTS, N. M., FOWLER, H. J., ROBERTS, M. J., CHAN, STEVEN, C. and SENIOR, C. A. (2014). Heavier summer downpours with climate change revealed by weather forecast resolution model. *Nat. Clim. Change* **4** 570–576.
- KENDON, E. J., FOSSER, G., MURPHY, J., CHAN, S., CLARK, R., HARRIS, G., LOCK, A., LOWE, J., MARTIN, G. et al. (2019). UKCP convection-permitting model projections: Science report.
- NADARAJAH, S., ANDERSON, C. W. and TAWN, J. A. (1998). Ordered multivariate extremes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 473–496. [MR1616073](#) <https://doi.org/10.1111/1467-9868.00136>
- NORTHROP, P. J. and JONATHAN, P. (2011). Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics* **22** 799–809. [MR2861046](#) <https://doi.org/10.1002/env.1106>
- MET OFFICE HADLEY CENTRE (2019). UKCP local projections on a 5 km grid over the UK for 1980–2080. Accessed 24/01/2022. Available at <https://catalogue.ceda.ac.uk/uuid/e304987739e04cdc960598fa5e4439d0>.
- 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>
- PALACIOS-RODRÍGUEZ, F., TOULEMONDE, G. and CARREAU, J. (2020). Generalized Pareto processes for simulating space-time extreme events: An application to precipitation reanalyses. *Stoch. Environ. Res. Risk Assess.* **34** 2033–2052.
- POLITIS, D. N. and ROMANO, J. P. (1994). The stationary bootstrap. *J. Amer. Statist. Assoc.* **89** 1303–1313. [MR1310224](#)
- REICH, B. J. and SHABY, B. A. (2012). A hierarchical max-stable spatial model for extreme precipitation. *Ann. Appl. Stat.* **6** 1430–1451. [MR3058670](#) <https://doi.org/10.1214/12-AOAS591>
- RICHARDS, J. and TAWN, J. A. (2022). On the tail behaviour of aggregated random variables. arXiv:e-prints. Available at [arXiv:2105.11917](https://arxiv.org/abs/2105.11917).
- RICHARDS, J., TAWN, J. A. and BROWN, S. (2022). Supplement to “Modelling extremes of spatial aggregates of precipitation using conditional methods.” <https://doi.org/10.1214/22-AOAS1609SUPPA>, <https://doi.org/10.1214/22-AOAS1609SUPPB>
- RICHARDS, J. and WADSWORTH, J. L. (2021). Spatial deformation for nonstationary extremal dependence. *Environmetrics* **32** Paper No. e2671, 22. [MR4307890](#) <https://doi.org/10.1002/env.2671>
- RODRIGUEZ-ITURBE, I., COX, D. R. and ISHAM, V. (1987). Some models for rainfall based on stochastic point processes. *Proc. R. Soc. Lond., Math. Phys. Sci.* **410** 269–288. [MR0887878](#)
- SCHROEER, K., KIRCHENGAST, G. and SUNGMIN, O. (2018). Strong dependence of extreme convective precipitation intensities on gauge network density. *Geophys. Res. Lett.* **45** 8253–8263.
- SHARKEY, P. and WINTER, H. C. (2019). A Bayesian spatial hierarchical model for extreme precipitation in Great Britain. *Environmetrics* **30** e2529, 19. [MR3908107](#) <https://doi.org/10.1002/env.2529>
- SHOOTER, R., TAWN, J., ROSS, E. and JONATHAN, P. (2021). Basin-wide spatial conditional extremes for severe ocean storms. *Extremes* **24** 241–265. [MR4246277](#) <https://doi.org/10.1007/s10687-020-00389-w>
- SIMPSON, E. S., OPITZ, T. and WADSWORTH, J. L. (2021). High-dimensional modeling of spatial and spatio-temporal conditional extremes using INLA and the SPDE approach. arXiv:e-prints. Available at [arXiv:2011.04486](https://arxiv.org/abs/2011.04486).

- SIMPSON, E. S. and WADSWORTH, J. L. (2021). Conditional modelling of spatio-temporal extremes for Red Sea surface temperatures. *Spat. Stat.* **41** 100482.
- THOMASSEN, E. D., SØRUP, H. J. D., SCHEIBEL, M., EINFALT, T. and ARNBJERG-NIELSEN, K. (2020). Data-driven distinction between convective, frontal and mixed extreme rainfall events in radar data. *Hydrol. Earth Syst. Sci. Discuss.* **2020** 1–26.
- WADSWORTH, J. L. and TAWN, J. A. (2019). Higher-dimensional spatial extremes via single-site conditioning. arXiv:e-prints. Available at [arXiv:1912.06560](https://arxiv.org/abs/1912.06560).
- WESTRA, S. and SISSON, S. A. (2011). Detection of non-stationarity in precipitation extremes using a max-stable process model. *J. Hydrol.* **406** 119–128.
- WOOD, S. N. (2006). *Generalized Additive Models: An Introduction with R. Texts in Statistical Science Series*. CRC Press/CRC, Boca Raton, FL. [MR2206355](https://doi.org/10.1080/01621459.2018.1529596)
- 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)
<https://doi.org/10.1080/01621459.2018.1529596>

A SPATIAL CAUSAL ANALYSIS OF WILDLAND FIRE-CONTRIBUTED PM_{2.5} USING NUMERICAL MODEL OUTPUT

BY ALEXANDRA LARSEN^{1,a}, SHU YANG^{2,b}, BRIAN J. REICH^{2,c} AND ANA G. RAPPOLD^{3,d}

¹Department of Biostatistics and Bioinformatics, Duke University, ^aalexandra.larsen@duke.edu

²Department of Statistics, North Carolina State University, ^bsyang24@ncsu.edu, ^cbjreich@ncsu.edu

³National Health and Environmental Effects Research Laboratory—Environmental Public Health Division, US Environmental Protection Agency, ^dRappold.Ana@epa.gov

Wildland fire smoke contains hazardous levels of fine particulate matter (PM_{2.5}), a pollutant shown to adversely effect health. Estimating fire attributable PM_{2.5} concentrations is key to quantifying the impact on air quality and subsequent health burden. This is a challenging problem since only total PM_{2.5} is measured at monitoring stations and both fire-attributable PM_{2.5} and PM_{2.5} from all other sources are correlated in space and time. We propose a framework for estimating fire-contributed PM_{2.5} and PM_{2.5} from all other sources using a novel causal inference framework and bias-adjusted chemical model representations of PM_{2.5} under counterfactual scenarios. The chemical model representation of PM_{2.5} for this analysis is simulated using Community Multiscale Air Quality Modeling System (CMAQ), run with and without fire emissions across the contiguous U.S. for the 2008–2012 wildfire seasons. The CMAQ output is calibrated with observations from monitoring sites for the same spatial domain and time period. We use a Bayesian model that accounts for spatial variation to estimate the effect of wildland fires on PM_{2.5} and state assumptions under which the estimate has a valid causal interpretation. Our results include estimates of the contributions of wildfire smoke to PM_{2.5} for the contiguous U.S. Additionally, we compute the health burden associated with the PM_{2.5} attributable to wildfire smoke.

REFERENCES

- ALLEN, M. R. and STOTT, P. A. (2003). Estimating signal amplitudes in optimal fingerprinting, Part I: Theory. *Clim. Dyn.* **21** 477–491. <https://doi.org/10.1007/s00382-003-0313-9>
- BENMAP (2017). Environmental Benefits Mapping and Analysis Program—Community Edition User’s Manual Appendices Technical Report.
- BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *J. Agric. Biol. Environ. Stat.* **15** 176–197. [MR2787270 https://doi.org/10.1007/s13253-009-0004-z](https://doi.org/10.1007/s13253-009-0004-z)
- CHILÈS, J.-P. and DELFINER, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*, 2nd ed. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ. [MR2850475 https://doi.org/10.1002/9781118136188](https://doi.org/10.1002/9781118136188)
- CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA. <https://doi.org/10.1002/9781119115151>
- DELFINO, R. J., BRUMMEL, S., WU, J., STERN, H., OSTRO, B., LIPSETT, M., WINER, A., STREET, D. H., ZHANG, L. et al. (2009). The relationship of respiratory and cardiovascular hospital admissions to the southern California wildfires of 2003. *Occup. Environ. Med.* **66** 189–197. <https://doi.org/10.1136/oem.2008.041376>
- DENNEKAMP, M. and ABRAMSON, M. J. (2011). The effects of bushfire smoke on respiratory health. *Respirology* **16** 198–209. <https://doi.org/10.1111/j.1440-1843.2010.01868.x>
- DENNEKAMP, M., STRANEY, L. D., ERBAS, B., ABRAMSON, M. J., KEYWOOD, M., SMITH, K., SIM, M. R., GLASS, D. C., DEL MONACO, A. et al. (2015). Forest fire smoke exposures and out-of-hospital cardiac arrests in Melbourne, Australia: A case-crossover study. *Environ. Health Perspect.* **123** 959–964. <https://doi.org/10.1289/ehp.1408436>

- DOMINICI, F., GREENSTONE, M. and SUNSTEIN, C. R. (2014). Science and regulation. Particulate matter matters. *Science* **344** 257–259. <https://doi.org/10.1126/science.1247348>
- FUENTES, M. and RAFTERY, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61** 36–45. MR2129199 <https://doi.org/10.1111/j.0006-341X.2005.030821.x>
- GELFAND, A. E., KIM, H.-J., SIRMANS, C. F. and BANERJEE, S. (2003). Spatial modeling with spatially varying coefficient processes. *J. Amer. Statist. Assoc.* **98** 387–396. MR1995715 <https://doi.org/10.1198/016214503000170>
- GELFAND, A. E., SCHMIDT, A. M., BANERJEE, S. and SIRMANS, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST* **13** 263–312. MR2154003 <https://doi.org/10.1007/BF02595775>
- HAIKERWAL, A., AKRAM, M., DEL MONACO, A., SMITH, K., SIM, M. R., MEYER, M., TONKIN, A. M., ABRAMSON, M. J. and DENNEKAMP, M. (2015). Impact of fine particulate matter (PM_{2.5}) exposure during wildfires on cardiovascular health outcomes. *J. Am. Heart Assoc.* **4** e001653. <https://doi.org/10.1161/JAHA.114.001653>
- HAIKERWAL, A., AKRAM, M., SIM, M. R., MEYER, M., ABRAMSON, M. J. and DENNEKAMP, M. (2016). Fine particulate matter (PM_{2.5}) exposure during a prolonged wildfire period and emergency department visits for asthma. *Respirology* **21** 88–94. <https://doi.org/10.1111/resp.12613>
- HALLORAN, M. E. and STRUCHINER, C. J. (1991). Study designs for dependent happenings. *Epidemiology* **2** 331–8.
- KATZFUSS, M., HAMMERLING, D. and SMITH, R. L. (2017). A Bayesian hierarchical model for climate change detection and attribution. *Geophys. Res. Lett.* **44** 5720–5728. <https://doi.org/10.1002/2017GL073688>
- HANNART, A., PEARL, J., OTTO, F. E. L., NAVEAU, P. and GHIL, M. (2016). Causal counterfactual theory for the attribution of weather and climate-related events. *Bull. Am. Meteorol. Soc.* **97** 99–110. <https://doi.org/10.1175/BAMS-D-14-00034.1>
- HANSEN, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika* **95** 481–488. MR2521594 <https://doi.org/10.1093/biomet/asn004>
- HEGERL, G. and ZWIERS, F. (2011). Use of models in detection and attribution of climate change. *Wiley Interdiscip. Rev.: Clim. Change* **2** 570–591. <https://doi.org/10.1002/wcc.121>
- HERNÁN, M. A., ALONSO, A., LOGAN, R., GRODSTEIN, F., MICHELS, K. B., WILLETT, W. C., MANSON, J. E. and ROBINS, J. M. (2008). Observational studies analyzed like randomized experiments. *Epidemiology* **19** 766–779. <https://doi.org/10.1097/EDE.0b013e3181875e61>
- HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618
- HONG, G. and RAUDENBUSH, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J. Amer. Statist. Assoc.* **101** 901–910. MR2324091 <https://doi.org/10.1198/016214506000000447>
- HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 <https://doi.org/10.1198/016214508000000292>
- JOHNSTON, F. H., HENDERSON, S. B., CHEN, Y., RANDERSON, J. T., MARLIER, M., DEFRIES, R. S., KINNEY, P., BOWMAN, D. M. J. S. and BRAUER, M. (2012). Estimated global mortality attributable to smoke from landscape fires. *Environ. Health Perspect.* **120** 695–701. <https://doi.org/10.1289/ehp.1104422>
- KAO, E. K. (2017). *Causal Inference Under Network Interference: A Framework for Experiments on Social Networks*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Harvard University. MR4172385
- 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>
- KNUTSON, T., KOSSIN, J. P., MEARS, C., PERLWITZ, J. and WEHNER, M. F. (2017). Detection and attribution of climate change. In *Climate Science Special Report: Fourth National Climate Assessment, Volume I* 114–132. U.S. Global Change Research Program, Washington, DC, USA. <https://doi.org/10.7930/J01834ND>
- LARSEN, A., YANG, S., REICH, B. J. and RAPPOLD, A. G. (2022). Supplement to “A spatial causal analysis of wildland fire-contributed PM_{2.5} using numerical model output—REVISED.” <https://doi.org/10.1214/22-AOAS1610SUPP>
- MCKENZIE, D., SHANKAR, U., KEANE, R. E., STAVROS, E. N., HEILMAN, W. E., FOX, D. G. and RIEBAU, A. C. (2014). Earth’s future smoke consequences of new wildfire regimes driven by climate change. *Earth’s Future* **2** 35–59. <https://doi.org/10.1002/2013EF000180>
- NATIONAL ACADEMIES OF SCIENCE (2016). Attribution of Extreme Weather Events in the Context of Climate Change Technical Report, National Academies of Sciences, Engineering, and Medicine, Washington, D.C. <https://doi.org/10.17226/21852>
- PAPADOGEORGOU, G., MEALLI, F. and ZIGLER, C. M. (2019). Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics* **75** 778–787. MR4012083 <https://doi.org/10.1111/biom.13049>

- RAPPOLD, A. G., STONE, S. L., CASCIO, W. E., NEAS, L. M., KILARU, V. J., CARRAWAY, M. S., SZYKMAN, J. J., ISING, A., CLEVE, W. E. et al. (2011). Peat bog wildfire smoke exposure in rural North Carolina is associated with cardiopulmonary emergency department visits assessed through syndromic surveillance. *Environ. Health Perspect.* **119** 1415–1420. <https://doi.org/10.1289/ehp.1003206>
- RAPPOLD, A. G., REYES, J., POULIOT, G., CASCIO, W. E. and DIAZ-SANCHEZ, D. (2017). Community vulnerability to health impacts of wildland fire smoke exposure. *Environ. Sci. Technol.* **51** 6674–6682. <https://doi.org/10.1021/acs.est.6b06200>
- REICH, B. J., YANG, S., GUAN, Y., GIFFIN, A. B., MILLER, M. J. and RAPPOLD, A. (2021). A review of spatial causal inference methods for environmental and epidemiological applications. *Int. Stat. Rev.* **89** 605–634. [MR4411920 https://doi.org/10.1111/insr.12452](https://doi.org/10.1111/insr.12452)
- ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *J. Amer. Statist. Assoc.* **102** 191–200. [MR2345537 https://doi.org/10.1198/016214506000001112](https://doi.org/10.1198/016214506000001112)
- RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. [MR0472152](https://doi.org/10.1214/aos/1176344972)
- SCHMIDT, A. M. and GELFAND, A. E. (2003). A Bayesian coregionalization approach for multivariate pollutant data. *J. Geophys. Res., Atmos.* **108**. <https://doi.org/10.1029/2002JD002905>
- US FOREST SERVICE (2015). PNW—Fire and Environmental Research Applications Team (FERA) Research/Studies.
- SOBEL, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Amer. Statist. Assoc.* **101** 1398–1407. [MR2307573 https://doi.org/10.1198/016214506000000636](https://doi.org/10.1198/016214506000000636)
- STAVROS, E. N., MCKENZIE, D. and LARKIN, N. (2014). The climate-wildfire-air quality system: Interactions and feedbacks across spatial and temporal scales. *Wiley Interdiscip. Rev.: Clim. Change* **5** 719–733. <https://doi.org/10.1002/wcc.303>
- TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21** 55–75. [MR2867538 https://doi.org/10.1177/0962280210386779](https://doi.org/10.1177/0962280210386779)
- USEPA (2010). Quantitative Health Risk Assessment for Particulate Matter Technical Report US Environmental Protection Agency.
- USEPA (2015). AirData Download Files Documentation.
- USEPA (2019). Overview of science processes in CMAQ.
- USEPA (2020). Particulate Matter (PM_{2.5}) Trends.
- WETTSTEIN, Z. S., HOSHIKO, S., FAHIMI, J., HARRISON, R. J., CASCIO, W. E. and RAPPOLD, A. G. (2018). Cardiovascular and cerebrovascular emergency department visits associated with wildfire smoke exposure in California in 2015. *J. Am. Heart Assoc.* **7** e007492. <https://doi.org/10.1161/JAHA.117.007492>
- ZIGLER, C. M., CHOIRAT, C. and DOMINICI, F. (2018). Impact of national ambient air quality standards nonattainment designations on particulate pollution and health. *Epidemiology* **29** 165–174. <https://doi.org/10.1097/EDE.0000000000000777>
- ZIGLER, C. M., DOMINICI, F. and WANG, Y. (2012). Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics* **13** 289–302. <https://doi.org/10.1093/biostatistics/kxr052>
- ZIGLER, C. M. and PAPADOGEORGOU, G. (2021). Bipartite causal inference with interference. *Statist. Sci.* **36** 109–123. [MR4194206 https://doi.org/10.1214/19-STS749](https://doi.org/10.1214/19-STS749)

DIMENSIONS, POWER AND FACTORS IN AN OBSERVATIONAL STUDY OF BEHAVIORAL PROBLEMS AFTER PHYSICAL ABUSE OF CHILDREN

BY TING YE^{1,a}, DYLAN S. SMALL^{2,b} AND PAUL R. ROSENBAUM^{2,c}

¹Department of Biostatistics, University of Washington, ^atingye1@uw.edu

²Department of Statistics and Data Science, Wharton School, University of Pennsylvania, ^bdsmall@wharton.upenn.edu,
^crosenbaum@wharton.upenn.edu

Many observational studies assess the impact of a treatment on an outcome that has several dimensions. In the observational study that we discuss, physical abuse of children may affect the degree to which the child exhibits depression, withdrawal or aggression. A treatment may affect all, some or none of these dimensions. In addition to the scientific interest in learning the effect on each dimension, it is also known that an appropriate combination of dimensions may increase power, efficiency and insensitivity to unmeasured biases; however, finding this appropriate combination requires corrections for multiple testing that erode power. We explore this trade-off by developing a new formula for the power of a sensitivity analysis in a simple situation with several dimensions. The methodology is applied to study the effects of physical abuse in early childhood and its possible effects on several dimensions of subsequent behavioral problems. Also, a general method is proposed for converting any signed rank test for matched pairs into an analogous test for matching each treated individual to several controls, and the performance of this extension is examined. The proposed method aids in studying the relative magnitude of the effect on different dimensions. A second evidence factor considers the dose or intensity of physical abuse.

REFERENCES

- ACHENBACH, T. M. and RESCORLA, L. A. (2000). *Manual for the ASEBA Preschool Forms and Profiles*. Univ. Vermont, Research Center for Children, Youth and Families, Burlington, VT.
- AMMERMAN, R. T., CASSISI, J. E., HERSEN, M. and VAN HASSELT, V. B. (1986). Consequences of physical abuse and neglect in children. *Clinical Psychol. Rev.* **6** 291–310.
- CHOI, J. K., WANG, D. and JACKSON, A. P. (2019). Adverse experiences in early childhood and their longitudinal impact on later behavioral problems of children living in poverty. *Child Abuse Negl.* **98** 104–181.
- DURRANT, J. and ENSOM, R. (2012). Physical punishment of children: Lessons from 20 years of research. *CMAJ, Can. Med. Assoc. J.* **184** 1373–1377.
- GASTWIRTH, J. L., KRIEGER, A. M. and ROSENBAUM, P. R. (2000). Asymptotic separability in sensitivity analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 545–555. [MR1772414 https://doi.org/10.1111/1467-9868.00249](https://doi.org/10.1111/1467-9868.00249)
- HETTMANSPERGER, T. P. and MCKEAN, J. W. (2011). *Robust Nonparametric Statistical Methods*, 2nd ed. *Monographs on Statistics and Applied Probability* **119**. CRC Press, Boca Raton, FL. [MR2779026](https://doi.org/10.1080/01621459.2012.742018)
- HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6** 65–70.
- 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](https://doi.org/10.1080/01621459.2012.742018)
- HUNT, T. K. A., SLACK, K. S. and BERGER, L. M. (2017). Adverse childhood experiences and behavioral problems in middle childhood. *Child Abuse Negl.* **67** 391–402. <https://doi.org/10.1016/j.chiabu.2016.11.005>
- KARMAKAR, B., FRENCH, B. and SMALL, D. S. (2019). Integrating the evidence from evidence factors in observational studies. *Biometrika* **106** 353–367. [MR3949308 https://doi.org/10.1093/biomet/asz003](https://doi.org/10.1093/biomet/asz003)
- KARMAKAR, B., SMALL, D. S. and ROSENBAUM, P. R. (2020). Using evidence factors to clarify exposure biomarkers. *Am. J. Epidemiol.* **189** 243–249.

Key words and phrases. Causal inference, coherence among multiple outcomes, design sensitivity, evidence factors, observational study, power of a sensitivity analysis, Scheffé correction, sensitivity analysis.

- LEE, S. J., ALTSCHUL, I. and GERSHOFF, E. T. (2015). Wait until your father gets home? Mother's and fathers' spanking and development of child aggression. *Child. Youth Serv. Rev.* **52** 158–166. <https://doi.org/10.1016/j.chilcyouth.2014.11.006>
- 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](https://doi.org/10.1111/biomet.12373)
- LEHMANN, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day Series in Probability and Statistics. Holden-Day, San Francisco, CA. [MR0395032](https://doi.org/10.1002/9781118160701)
- MACKENZIE, M., NICKLAS, E., BROOKS-GUNN, J. and WALDFOGEL, J. (2015). Spanking and children's externalizing behavior across the first decade of life: Evidence for transactional processes. *Journal of Youth and Adolescence* **44** 658–669.
- MCKILLIP, J. (1992). Research without control groups: A control construct design. In *Methodological Issues in Applied Social Psychology* (F. B. Bryant et al., eds.) 159–175. Plenum, NY.
- NEYMAN, J. (1923, 1990). On the application of probability theory to agricultural experiments. *Statist. Sci.* 465–472.
- NOETHER, G. E. (1967). *Elements of Nonparametric Statistics*. Wiley, New York. [MR0219184](https://doi.org/10.1002/9781118160701)
- ORNDUFF, S. R. (2000). Childhood maltreatment and malevolence: Quantitative research findings. *Clin Psychol Rev* **20** 997–1018. [https://doi.org/10.1016/s0272-7358\(99\)00021-5](https://doi.org/10.1016/s0272-7358(99)00021-5)
- OSTERMAN, K., BJÖRKQVIST, K. and WAHLBECK, K. (2014). Twenty-eight years after the complete ban on the physical punishment of children in Finland: Trends and psychosocial concomitants. *Aggressive Behavior* **40** 568–581.
- PRATT, J. W. and GIBBONS, J. D. (2012). *Concepts of Nonparametric Theory*. Springer Series in Statistics. Springer, New York.
- RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley Series in Probability and Mathematical Statistics. Wiley, New York. [MR0346957](https://doi.org/10.1002/9781118160701)
- ROSENBAUM, P. R. (1997). Signed rank statistics for coherent predictions. *Biometrics* 556–566.
- ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. Springer Series in Statistics. Springer, New York. [MR1899138](https://doi.org/10.1007/978-1-4757-3692-2) <https://doi.org/10.1007/978-1-4757-3692-2>
- ROSENBAUM, P. R. (2004). Design sensitivity in observational studies. *Biometrika* **91** 153–164. [MR2050466](https://doi.org/10.1093/biomet/91.1.153) <https://doi.org/10.1093/biomet/91.1.153>
- ROSENBAUM, P. R. (2010). Evidence factors in observational studies. *Biometrika* **97** 333–345. [MR2650742](https://doi.org/10.1093/biomet/asq019) <https://doi.org/10.1093/biomet/asq019>
- ROSENBAUM, P. R. (2014). Weighted *M*-statistics with superior design sensitivity in matched observational studies with multiple controls. *J. Amer. Statist. Assoc.* **109** 1145–1158. [MR3265687](https://doi.org/10.1080/01621459.2013.879261) <https://doi.org/10.1080/01621459.2013.879261>
- ROSENBAUM, P. R. (2016a). Using Scheffé projections for multiple outcomes in an observational study of smoking and periodontal disease. *Ann. Appl. Stat.* **10** 1447–1471. [MR3553231](https://doi.org/10.1214/16-AOAS942) <https://doi.org/10.1214/16-AOAS942>
- ROSENBAUM, P. R. (2016b). The cross-cut statistic and its sensitivity to bias in observational studies with ordered doses of treatment. *Biometrics* **72** 175–183. [MR3500586](https://doi.org/10.1111/biom.12373) <https://doi.org/10.1111/biom.12373>
- ROSENBAUM, P. R. (2017). The general structure of evidence factors in observational studies. *Statist. Sci.* **32** 514–530. [MR3730520](https://doi.org/10.1214/17-STS621) <https://doi.org/10.1214/17-STS621>
- ROSENBAUM, P. R. (2020a). Combining planned and discovered comparisons in observational studies. *Biostatistics* **21** 384–399. [MR4120329](https://doi.org/10.1093/biostatistics/kxy055) <https://doi.org/10.1093/biostatistics/kxy055>
- ROSENBAUM, P. R. (2020b). *Design of Observational Studies*, 2nd ed. Springer Series in Statistics. Springer, New York. [MR2561612](https://doi.org/10.1007/978-1-4419-1213-8) <https://doi.org/10.1007/978-1-4419-1213-8>
- ROSENBAUM, P. R. (2020c). A conditional test with demonstrated insensitivity to unmeasured bias in matched observational studies. *Biometrika* **107** 827–840. [MR4186490](https://doi.org/10.1093/biomet/asaa032) <https://doi.org/10.1093/biomet/asaa032>
- ROSENBAUM, P. R. (2021). *Replication and Evidence Factors in Observational Studies*. CRC Press/CRC, NY.
- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.
- SCHEFFÉ, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika* **40** 87–104. [MR0057504](https://doi.org/10.2307/2333100) <https://doi.org/10.2307/2333100>
- SEN, P. K. (1968). On a further robustness property of the test and estimator based on Wilcoxon's signed rank statistic. *Ann. Math. Stat.* **39** 282–285. [MR0221697](https://doi.org/10.1214/aoms/1177698535) <https://doi.org/10.1214/aoms/1177698535>
- SILBER, J. H., ROSENBAUM, P. R., MCHUGH, M. D., LUDWIG, J. M., SMITH, H. L., NIKNAM, B. A., EVEN-SHOSHAN, O., FLEISHER, L. A., KELZ, R. R. et al. (2016). Comparison of the value of nursing work environments in hospitals across different levels of patient risk. *JAMA Surg.* **151** 527–536.
- TCHETGEN TCHETGEN, E. J. (2013). The control outcome calibration approach for causal inference with unobserved confounding. *Am. J. Epidemiol.* **179** 633–640.

- WARD, K. P., LEE, S., PACE, G., GROGAN-KAYLOR, A. and MA, J. (2020). Attachment style and the association of spanking and child externalizing behavior. *Academic Pediatrics* **20** 501–507.
- YE, T., SMALL, D. S. and ROSENBAUM, P. S. (2022). Supplement to “Dimensions, power and factors in an observational study of behavioral problems after physical abuse of children.” <https://doi.org/10.1214/22-AOAS1611SUPPA>, <https://doi.org/10.1214/22-AOAS1611SUPPB>
- ZAYKIN, D. V., ZHIVOTOVSKY, L. A., WESTFALL, P. H. and WEIR, B. S. (2002). Truncated product method for combining P-values. *Genet. Epidemiol.* **22** 170–185.
- ZOLOTOR, A. J., THEODORE, A. D., CHANG, J. J., BERKOFF, M. C. and RUNYAN, D. K. (2008). Speak softly—and forget the stick: Corporal punishment and child physical abuse. *Am. J. Prev. Med.* **35** 364–369.

SCALAR ON NETWORK REGRESSION VIA BOOSTING

BY EMILY L. MORRIS^a, KEVIN HE^b AND JIAN KANG^c

Department of Biostatistics, University of Michigan, ^aemorrisl@umich.edu, ^bkevinhe@umich.edu, ^cjiankang@umich.edu

Neuroimaging studies have a growing interest in learning the association between the individual brain connectivity networks and their clinical characteristics. It is also of great interest to identify the sub-brain networks as biomarkers to predict the clinical symptoms, such as disease status, potentially providing insight on neuropathology. This motivates the need for developing a new type of regression model where the response variable is scalar, and predictors are networks that are typically represented as adjacent matrices or weighted adjacent matrices to which we refer as scalar-on-network regression. In this work we develop a new boosting method for model fitting with subnetwork markers selection. Our approach, as opposed to group lasso or other existing regularization methods, is, essentially, a gradient descent algorithm leveraging known network structure. We demonstrate the utility of our methods via simulation studies and analysis of the resting-state fMRI data in a cognitive developmental cohort study.

REFERENCES

- ARROYO RELIÓN, J. D., KESSLER, D., LEVINA, E. and TAYLOR, S. F. (2019). Network classification with applications to brain connectomics. *Ann. Appl. Stat.* **13** 1648–1677. [MR4019153](#) <https://doi.org/10.1214/19-AOAS1252>
- BÜHLMANN, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* **34** 559–583. [MR2281878](#) <https://doi.org/10.1214/009053606000000092>
- BÜHLMANN, P. and YU, B. (2003). Boosting with the L_2 loss: Regression and classification. *J. Amer. Statist. Assoc.* **98** 324–339. [MR1995709](#) <https://doi.org/10.1198/0162145030000125>
- CAVANNA, A. E. (2007). The precuneus and consciousness. *CNS Spectr.* **12** 545–552.
- CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. [MR2443189](#) <https://doi.org/10.1093/biomet/asn034>
- CHEN, Y., WANG, X., KONG, L. and ZHU, H. (2016). Local region sparse learning for image-on-scalar regression. Preprint [arXiv:1605.08501](https://arxiv.org/abs/1605.08501).
- DESIKAN, R. S., SÉGONNE, F., FISCHL, B., QUINN, B. T., DICKERSON, B. C., BLACKER, D., BUCKNER, R. L., DALE, A. M., MAGUIRE, R. P. et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31** 968–980.
- EICKHOFF, S. B., YEO, B. T. and GENON, S. (2018). Imaging-based parcellations of the human brain. *Nat. Rev., Neurosci.* **19** 672–686.
- ERHARDT, E. B., RACHAKONDA, S., BEDRICK, E. J., ALLEN, E. A., ADALI, T. and CALHOUN, V. D. (2011). Comparison of multi-subject ICA methods for analysis of fMRI data. *Hum. Brain Mapp.* **32** 2075–2095.
- FOX, M. D. and GREICIUS, M. (2010). Clinical applications of resting state functional connectivity. *Front. Syst. Neurosci.* **4** 19.
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.
- FRITH, C. D. and FRITH, U. (1999). Interacting minds—a biological basis. *Science* **286** 1692–1695.
- FURL, N., GARRIDO, L., DOLAN, R. J., DRIVER, J. and DUCHAINE, B. (2011). Fusiform gyrus face selectivity relates to individual differences in facial recognition ability. *J. Cogn. Neurosci.* **23** 1723–1740.
- GOLDSMITH, J., HUANG, L. and CRAINICEANU, C. M. (2014). Smooth scalar-on-image regression via spatial Bayesian variable selection. *J. Comput. Graph. Statist.* **23** 46–64. [MR3173760](#) <https://doi.org/10.1080/10618600.2012.743437>
- GONZÁLEZ-RECIO, O., WEIGEL, K. A., GIANOLA, D., NAYA, H. and ROSA, G. J. (2010). L 2-boosting algorithm applied to high-dimensional problems in genomic selection. *Genet. Research* **92** 227–237.

- GORDON, E. M., LAUMANN, T. O., ADEYEMO, B., HUCKINS, J. F., KELLEY, W. M. and PETERSEN, S. E. (2016). Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cereb. Cortex* **26** 288–303. <https://doi.org/10.1093/cercor/bhu239>
- GORDON, E. M., LAUMANN, T. O., MAREK, S., RAUT, R. V., GRATTON, C., NEWBOLD, D. J., GREENE, D. J., COALSON, R. S., SNYDER, A. Z. et al. (2020). Default-mode network streams for coupling to language and control systems. *Proc. Natl. Acad. Sci. USA* **117** 17308–17319.
- GREICIUS, M. D., FLORES, B. H., MENON, V., GLOVER, G. H., SOLVASON, H. B., KENNA, H., REISS, A. L. and SCHATZBERG, A. F. (2007). Resting-state functional connectivity in major depression: Abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biol. Psychiatry* **62** 429–437.
- GUHA, S. and RODRIGUEZ, A. (2020). High dimensional Bayesian network classification with network global-local shrinkage priors. Preprint [arXiv:2009.11401](https://arxiv.org/abs/2009.11401).
- GUHA, S. and RODRIGUEZ, A. (2021). Bayesian regression with undirected network predictors with an application to brain connectome data. *J. Amer. Statist. Assoc.* **116** 581–593. [MR4270005 https://doi.org/10.1080/01621459.2020.1772079](https://doi.org/10.1080/01621459.2020.1772079)
- KANG, J., REICH, B. J. and STAIKU, A.-M. (2018). Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika* **105** 165–184. [MR3768872 https://doi.org/10.1093/biomet/asx075](https://doi.org/10.1093/biomet/asx075)
- LI, X., WANG, L. and WANG, H. J. (2021). Sparse learning and structure identification for ultrahigh-dimensional image-on-scalar regression. *J. Amer. Statist. Assoc.* **116** 1994–2008. [MR4353728 https://doi.org/10.1080/01621459.2020.1753523](https://doi.org/10.1080/01621459.2020.1753523)
- LI, F., ZHANG, T., WANG, Q., GONZALEZ, M. Z., MARESH, E. L. and COAN, J. A. (2015). Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *Ann. Appl. Stat.* **9** 687–713. [MR3371331 https://doi.org/10.1214/15-AOAS818](https://doi.org/10.1214/15-AOAS818)
- LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. [MR2563983](https://doi.org/10.26434/chemrxiv-2012-09)
- MAYR, A., BINDER, H., GESELLER, O. and SCHMID, M. (2014). The evolution of boosting algorithms—from machine learning to statistical modelling. Preprint [arXiv:1403.1452](https://arxiv.org/abs/1403.1452).
- MORRIS, E. L., HE, K. and KANG, J. (2022). Supplement to “Scalar on network regression via boosting.” <https://doi.org/10.1214/22-AOAS1612SUPPA>, <https://doi.org/10.1214/22-AOAS1612SUPPB>
- PALEJWALA, A. H., O’CONNOR, K. P., MILTON, C. K., ANDERSON, C., PELARGOS, P., BRIGGS, R. G., CONNER, A. K., O’DONOGHUE, D. L., GLENN, C. A. et al. (2020). Anatomy and white matter connections of the fusiform gyrus. *Sci. Rep.* **10** 1–12.
- PATEL, R. S., BOWMAN, F. D. and RILLING, J. K. (2006). A Bayesian approach to determining connectivity of the human brain. *Hum. Brain Mapp.* **27** 267–276.
- POURTOIS, G., DE GELDER, B., BOL, A. and CROMMELINCK, M. (2005). Perception of facial expressions and voices and of their combination in the human brain. *Cortex* **41** 49–59. [https://doi.org/10.1016/s0010-9452\(08\)70177-1](https://doi.org/10.1016/s0010-9452(08)70177-1)
- POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M. et al. (2011a). Functional network organization of the human brain. *Neuron* **72** 665–678. <https://doi.org/10.1016/j.neuron.2011.09.006>
- POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M. et al. (2011b). Functional network organization of the human brain. *Neuron* **72** 665–678.
- RIÈS, S. K., DRONKERS, N. F. and KNIGHT, R. T. (2016). Choosing words: Left hemisphere, right hemisphere, or both? Perspective on the lateralization of word retrieval. *Ann. N.Y. Acad. Sci.* **1369** 111–131. <https://doi.org/10.1111/nyas.12993>
- SCHMIDT, A., SMIESKOVA, R., ASTON, J., SIMON, A., ALLEN, P., FUSAR-POLI, P., MCGUIRE, P. K., RIECHER-RÖSSLER, A., STEPHAN, K. E. et al. (2013). Brain connectivity abnormalities predating the onset of psychosis: Correlation with the effect of medication. *JAMA Psychiatr.* **70** 903–912.
- SMITH, S. M., FOX, P. T., MILLER, K. L., GLAHN, D. C., FOX, P. M., MACKAY, C. E., FILIPPINI, N., WATKINS, K. E., TORO, R. et al. (2009). Correspondence of the brain’s functional architecture during activation and rest. *Proc. Natl. Acad. Sci. USA* **106** 13040–13045.
- SRIKADA, C., RUTHERFORD, S., ANGSTADT, M., THOMPSON, W. K., LUCIANA, M., WEIGARD, A., HYDE, L. H. and HEITZEG, M. (2019). Prediction of neurocognition in youth from resting state fMRI. *Mol. Psychiatry* 1–9.
- SUPEKAR, K., MENON, V., RUBIN, D., MUSEN, M. and GREICIUS, M. D. (2008). Network analysis of intrinsic functional brain connectivity in Alzheimer’s disease. *PLoS Comput. Biol.* **4** e1000100. <https://doi.org/10.1371/journal.pcbi.1000100>
- THOMPSON, W. K., BARCH, D. M., BJORK, J. M., GONZALEZ, R., NAGEL, B. J., NIXON, S. J. and LUCIANA, M. (2019). The structure of cognition in 9 and 10 year-old children and associations with problem

- behaviors: Findings from the ABCD study's baseline neurocognitive battery. *Dev. Cogn. Neurosci.* **36** 100606. <https://doi.org/10.1016/j.dcn.2018.12.004>
- TZOURIO-MAZOYER, N., LANDEAU, B., PAPATHANASSIOU, D., CRIVELLO, F., ETARD, O., DELCROIX, N., MAZOYER, B. and JOLIOT, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15** 273–289.
- VOGELEY, K., MAY, M., RITZL, A., FALKAI, P., ZILLES, K. and FINK, G. R. (2004). Neural correlates of first-person perspective as one constituent of human self-consciousness. *J. Cogn. Neurosci.* **16** 817–827.
- WANG, L., ZHANG, Z. and DUNSON, D. (2019). Symmetric bilinear regression for signal subgraph estimation. *IEEE Trans. Signal Process.* **67** 1929–1940. [MR3960460 https://doi.org/10.1109/TSP.2019.2899818](https://doi.org/10.1109/TSP.2019.2899818)
- WANG, X. and ZHU, H. (2017). Generalized scalar-on-image regression models via total variation. *J. Amer. Statist. Assoc.* **112** 1156–1168. [MR3735367 https://doi.org/10.1080/01621459.2016.1194846](https://doi.org/10.1080/01621459.2016.1194846)
- WANG, Y., KANG, J., KEMMER, P. B. and GUO, Y. (2016). An efficient and reliable statistical method for estimating functional connectivity in large scale brain networks using partial correlation. *Front. Neurosci.* **10** 123.
- XIA, C. H., MA, Z., CIRIC, R., GU, S., BETZEL, R. F., KACZKURKIN, A. N., CALKINS, M. E., COOK, P. A., DE LA GARZA, A. G. et al. (2018). Linked dimensions of psychopathology and connectivity in functional brain networks. *Nat. Commun.* **9** 1–14.
- YANG, Y., ZOU, H. and BHATNAGAR, S. (2020). gglasso: Group Lasso penalized learning using a unified BMD algorithm. R package version 1.5.
- YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. [MR2212574 https://doi.org/10.1111/j.1467-9868.2005.00532.x](https://doi.org/10.1111/j.1467-9868.2005.00532.x)
- ZHANG, J., SUN, W. W. and LI, L. (2018). Network response regression for modeling population of networks with covariates. Preprint [arXiv:1810.03192](https://arxiv.org/abs/1810.03192).
- ZHANG, D., LI, L., SRIPADA, C. and KANG, J. (2020). Image-on-scalar regression via deep neural networks. Preprint [arXiv:2006.09911](https://arxiv.org/abs/2006.09911).
- ZHOU, H., LI, L. and ZHU, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Amer. Statist. Assoc.* **108** 540–552. [MR3174640 https://doi.org/10.1080/01621459.2013.776499](https://doi.org/10.1080/01621459.2013.776499)
- ZHU, H., FAN, J. and KONG, L. (2014). Spatially varying coefficient model for neuroimaging data with jump discontinuities. *J. Amer. Statist. Assoc.* **109** 1084–1098. [MR3265682 https://doi.org/10.1080/01621459.2014.881742](https://doi.org/10.1080/01621459.2014.881742)

MULTISCALE SPECTRAL MODELLING FOR NONSTATIONARY TIME SERIES WITHIN AN ORDERED MULTIPLE-TRIAL EXPERIMENT

BY JONATHAN EMBLETON^{1,a}, MARINA I. KNIGHT^{1,b} AND HERNANDO OMBAO^{2,c}

¹Department of Mathematics, University of York, a.je687@york.ac.uk, b.marina.knight@york.ac.uk

²Statistics Program, King Abdullah University of Science and Technology (KAUST), c.hernando.ombao@kaust.edu.sa

Within the neurosciences it is natural to observe variability across time in the dynamics of an underlying brain process. Wavelets are essential in analysing brain signals because, even within a single trial, brain signals exhibit nonstationary behaviour. However, neurological signals generated within an experiment may also potentially exhibit evolution across trials (replicates), even for identical stimuli. As neurologists consider localised spectra of brain signals to be most informative, we propose the MULTiple-Trials Locally Stationary Wavelet process (MULT-LSW) that fills the gap in the literature by directly giving a stochastic wavelet representation of the time series of ordered replicates itself. MULT-LSW yields a natural desired time- and trial-localisation of the process dynamics, capturing nonstationary behaviour both within and across trials. While current techniques are restricted by the assumption of uncorrelated replicates, here we account for between-trial correlation. We rigorously develop the associated wavelet spectral estimation framework along with its asymptotic properties. By means of thorough simulation studies, we demonstrate the theoretical estimator properties hold in practice. A real data investigation into the evolutionary dynamics of the hippocampus and nucleus accumbens, during an associative learning experiment, demonstrates the applicability of our proposed methodology as well as the new insights it provides. Our model is general and facilitates wider experimental data analysis than the current literature allows.

REFERENCES

- ABELA, A., DUAN, Y. and CHUDASAMA, Y. (2015). Hippocampal interplay with the nucleus accumbens is critical for decisions about time. *Eur. J. Neurosci.* **42** 2224–2233.
- ARIELI, A., STERKIN, A., GRINVALD, A. and AERTSEN, A. (1996). Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses. *Science* **273** 1868–1871.
- BUZSÁKI, G. (2006). *Rhythms of the Brain*. Oxford Univ. Press, Oxford. MR2270828 <https://doi.org/10.1093/acprof:oso/9780195301069.001.0001>
- CHAU, J. and VON SACHS, R. (2016). Functional mixed effects wavelet estimation for spectra of replicated time series. *Electron. J. Stat.* **10** 2461–2510. MR3545466 <https://doi.org/10.1214/16-EJS1181>
- CRYER, J. D. and CHAN, K-S. (2008). *Time Series Analysis with Applications in R*. Springer, Berlin.
- DAHLHAUS, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25** 1–37. MR1429916 <https://doi.org/10.1214/aos/1034276620>
- DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics **61**. SIAM, Philadelphia, PA. MR1162107 <https://doi.org/10.1137/1.9781611970104>
- DIGGLE, P. J. and AL WASEL, I. (1997). Spectral analysis of replicated biomedical time series. *J. R. Stat. Soc. Ser. C.* **46** 31–71. MR1452286 <https://doi.org/10.1111/1467-9876.00047>
- EMBLETON, J., KNIGHT, M. I. and OMBAO, H. (2020). Multiscale modelling of replicated nonstationary time series. Available at [arXiv:2005.09440](https://arxiv.org/abs/2005.09440).
- EMBLETON, J., KNIGHT, M. I. and OMBAO, H. (2022). Supplement to “Multiscale spectral modelling for nonstationary time series within an ordered multiple-trial experiment.” <https://doi.org/10.1214/22-AOAS1614SUPP>
- FIECAS, M. and OMBAO, H. (2016). Modeling the evolution of dynamic brain processes during an associative learning experiment. *J. Amer. Statist. Assoc.* **111** 1440–1453. MR3601700 <https://doi.org/10.1080/01621459.2016.1165683>

- FRYZLEWICZ, P. (2005). Modelling and forecasting financial log-returns as locally stationary wavelet processes. *J. Appl. Stat.* **32** 503–528. [MR2112368](#)
- FRYZLEWICZ, P. and NASON, G. P. (2006). Haar–Fisz estimation of evolutionary wavelet spectra. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 611–634. [MR2301011](#) <https://doi.org/10.1111/j.1467-9868.2006.00558.x>
- GORROSTIETA, C., OMBAO, H., PRADO, R., PATEL, S. and ESKANDAR, E. (2012). Exploring dependence between brain signals in a monkey during learning. *J. Time Series Anal.* **33** 771–778. [MR2969910](#) <https://doi.org/10.1111/j.1467-9892.2011.00767.x>
- GOTT, A. N., ECKLEY, I. A. and ASTON, J. A. D. (2015). Estimating the population local wavelet spectrum with application to non-stationary functional magnetic resonance imaging time series. *Stat. Med.* **34** 3901–3915. [MR3422153](#) <https://doi.org/10.1002/sim.6592>
- HARGREAVES, J. K., KNIGHT, M. I., PITCHFORD, J. W., OAKENFULL, R. J. and DAVIS, S. J. (2018). Clustering nonstationary circadian rhythms using locally stationary wavelet representations. *Multiscale Model. Simul.* **16** 184–214. [MR3749378](#) <https://doi.org/10.1137/16M1108078>
- HARGREAVES, J. K., KNIGHT, M. I., PITCHFORD, J. W., OAKENFULL, R. J., CHAWLA, S., MUNNS, J. and DAVIS, S. J. (2019). Wavelet spectral testing: Application to nonstationary circadian rhythms. *Ann. Appl. Stat.* **13** 1817–1846. [MR4019159](#) <https://doi.org/10.1214/19-AOAS1246>
- HOLLERMAN, J. and SCHULTZ, W. (1998). Dopamine neurons report an error in temporal prediction of reward during learning. *Nat. Neurosci.* **1** 304–309.
- HUK, A., BONNEN, K. and HE, B. J. (2018). Beyond trial-based paradigms: Continuous behaviour, ongoing neural activity, and natural stimuli. *J. Neurosci.* **38** 7551–7558.
- IACCARINO, H., SINGER, A. C., MARTORELL, A. J., RUDENKO, A., GAO, F., GILLINGHAM, T. Z., MATHYS, H., SEO, J., KRITSKIY, O. et al. (2016). Gamma frequency entrainment attenuates amyloid load and modifies microglia. *Nature* **540** 230–235.
- KILLICK, R., KNIGHT, M. I., NASON, G. P. and ECKLEY, I. A. (2020). The local partial autocorrelation function and some applications. *Electron. J. Stat.* **14** 3268–3314. [MR4147585](#) <https://doi.org/10.1214/20-EJS1748>
- KRAFTY, R. T., HALL, M. and GUO, W. (2011). Functional mixed effects spectral analysis. *Biometrika* **98** 583–598. [MR2836408](#) <https://doi.org/10.1093/biomet/asr032>
- MARTINEZ, J. G., BOHN, K. M., CARROLL, R. J. and MORRIS, J. S. (2013). A study of Mexican free-tailed bat chirp syllables: Bayesian functional mixed models for nonstationary acoustic time series. *J. Amer. Statist. Assoc.* **108** 514–526. [MR3174638](#) <https://doi.org/10.1080/01621459.2013.793118>
- MORRIS, J. S. (2015). Functional regression. *Annu. Rev. Stat. Appl.* **2** 321–359.
- MULDER, A. B., SHIBATA, R., TRULLIER, O. and WIENER, S. I. (2005). Spatially selective reward site responses in tonically active neurons of the nucleus accumbens in behaving rats. *Exp. Brain Res.* **163** 32–43. <https://doi.org/10.1007/s00221-004-2135-3>
- NASON, G. P. (2008). *Wavelet Methods in Statistics with R. Use R!* Springer, New York. [MR2445580](#) <https://doi.org/10.1007/978-0-387-75961-6>
- NASON, G. (2013). A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 879–904. [MR3124795](#) <https://doi.org/10.1111/rssb.12015>
- NASON, G. P. and SILVERMAN, B. W. (1994). The discrete wavelet transform in S. *J. Comput. Graph. Statist.* **3** 163–191.
- NASON, G. P., VON SACHS, R. and KROISANDT, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 271–292. [MR1749539](#) <https://doi.org/10.1111/1467-9868.00231>
- NEUMANN, M. H. and VON SACHS, R. (1997). Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Ann. Statist.* **25** 38–76. [MR1429917](#) <https://doi.org/10.1214/aos/1034276621>
- OH, H.-S., AMMANN, C. M., NAVEAU, P., NYCHKA, D. and OTTO-BLIESNER, B. L. (2003). Multi-resolution time series analysis applied to solar irradiance and climate reconstructions. *J. Atmos. Sol.-Terr. Phys.* **65** 191–201.
- OMBAO, H., VON SACHS, R. and GUO, W. (2005). SLEX analysis of multivariate nonstationary time series. *J. Amer. Statist. Assoc.* **100** 519–531. [MR2160556](#) <https://doi.org/10.1198/016214504000001448>
- OMBAO, H., RAZ, J., VON SACHS, R. and GUO, W. (2002). The SLEX model of a non-stationary random process. *Ann. Inst. Statist. Math.* **54** 171–200. [MR1893549](#) <https://doi.org/10.1023/A:1016130108440>
- PARK, T., ECKLEY, I. A. and OMBAO, H. C. (2014). Estimating time-evolving partial coherence between signals via multivariate locally stationary wavelet processes. *IEEE Trans. Signal Process.* **62** 5240–5250. [MR3268108](#) <https://doi.org/10.1109/TSP.2014.2343937>
- PRIESTLEY, M. B. (1965). Evolutionary spectra and non-stationary processes. *J. Roy. Statist. Soc. Ser. B* **27** 204–237. [MR0199886](#)

- QIN, L., GUO, W. and LITT, B. (2009). A time-frequency functional model for locally stationary time series data. *J. Comput. Graph. Statist.* **18** 675–693. [MR2572632](#) <https://doi.org/10.1198/jcgs.2009.06109>
- SAAVEDRA, P., HERNÁNDEZ, C. N. and ARTILES, J. (2000). Spectral analysis with replicated time series. *Comm. Statist. Theory Methods* **29** 2243–2362. [MR1802848](#) <https://doi.org/10.1080/03610920008832610>
- SANDERSON, J., FRYZLEWICZ, P. and JONES, M. W. (2010). Estimating linear dependence between non-stationary time series using the locally stationary wavelet model. *Biometrika* **97** 435–446. [MR2650749](#) <https://doi.org/10.1093/biomet/asq007>
- SCHULTZ, W., APICELLA, P., SCARNATI, E. and LJUNGBERG, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *J. Neurosci.* **12** 4595–4610.
- SEGER, C. and CINCOTTA, C. (2006). Dynamics of frontal, striatal, and hippocampal systems during rule learning. *Cereb. Cortex* **16** 1546–1555.
- TING, C.-M., OMBAO, H., SAMDIN, S. B. and SALLEH, S.-H. (2018). Estimating dynamic connectivity states in fMRI using regime-switching factor models. *IEEE Trans. Med. Imag.* **37** 1011–1023. <https://doi.org/10.1109/TMI.2017.2780185>
- VAN BELLEGEM, S. and VON SACHS, R. (2008). Locally adaptive estimation of evolutionary wavelet spectra. *Ann. Statist.* **36** 1879–1924. [MR2435459](#) <https://doi.org/10.1214/07-AOS524>
- VIDAKOVIC, B. (1999). *Statistical Modeling by Wavelets*. *Wiley Series in Probability and Statistics: Applied Probability and Statistics*. Wiley, New York. [MR1681904](#) <https://doi.org/10.1002/9780470317020>
- VON SACHS, R. and SCHNEIDER, K. (1996). Wavelet smoothing of evolutionary spectra by nonlinear thresholding. *Appl. Comput. Harmon. Anal.* **3** 268–282. [MR1400084](#) <https://doi.org/10.1006/acha.1996.0021>
- WIRTH, S., YANIKE, M., FRANK, L., SMITH, A., BROWN, E. and SUZUKI, W. (2003). Single neurons in the monkey hippocampus and learning of new associations. *Science* **300** 1578–1581.

BAYESIAN BI-CLUSTERING METHODS WITH APPLICATIONS IN COMPUTATIONAL BIOLOGY

BY HAN YAN^{1,a}, JIEXING WU^{2,c}, YANG LI^{3,d} AND JUN S. LIU^{1,b}

¹Department of Statistics, Harvard University, ^ahan_yan@g.harvard.edu, ^bjlui@stat.harvard.edu

²Google, ^cwujiexingharvard@gmail.com

³Citadel Securities, ^dyangli.stat@gmail.com

Bi-clustering is a useful approach in analyzing large biological data sets when the observations come from heterogeneous groups and have a large number of features. We outline a general Bayesian approach in tackling bi-clustering problems in moderate to high dimensions and propose three Bayesian bi-clustering models on categorical data which increase in complexities in their modeling of the distributions of features across bi-clusters. Our proposed methods apply to a wide range of scenarios: from situations where data are cluster-distinguishable only among a small subset of features but masked by a large amount of noise to situations where different groups of data are identified by different sets of features or data exhibit hierarchical structures. Through simulation studies we show that our methods outperform existing (bi-)clustering methods in both identifying clusters and recovering feature distributional patterns across bi-clusters. We further apply the developed approaches to a human genetic dataset, a human single-cell genomic dataset, and a collection of 1774 mouse genomic datasets with a focus on 58 genes from two pathways.

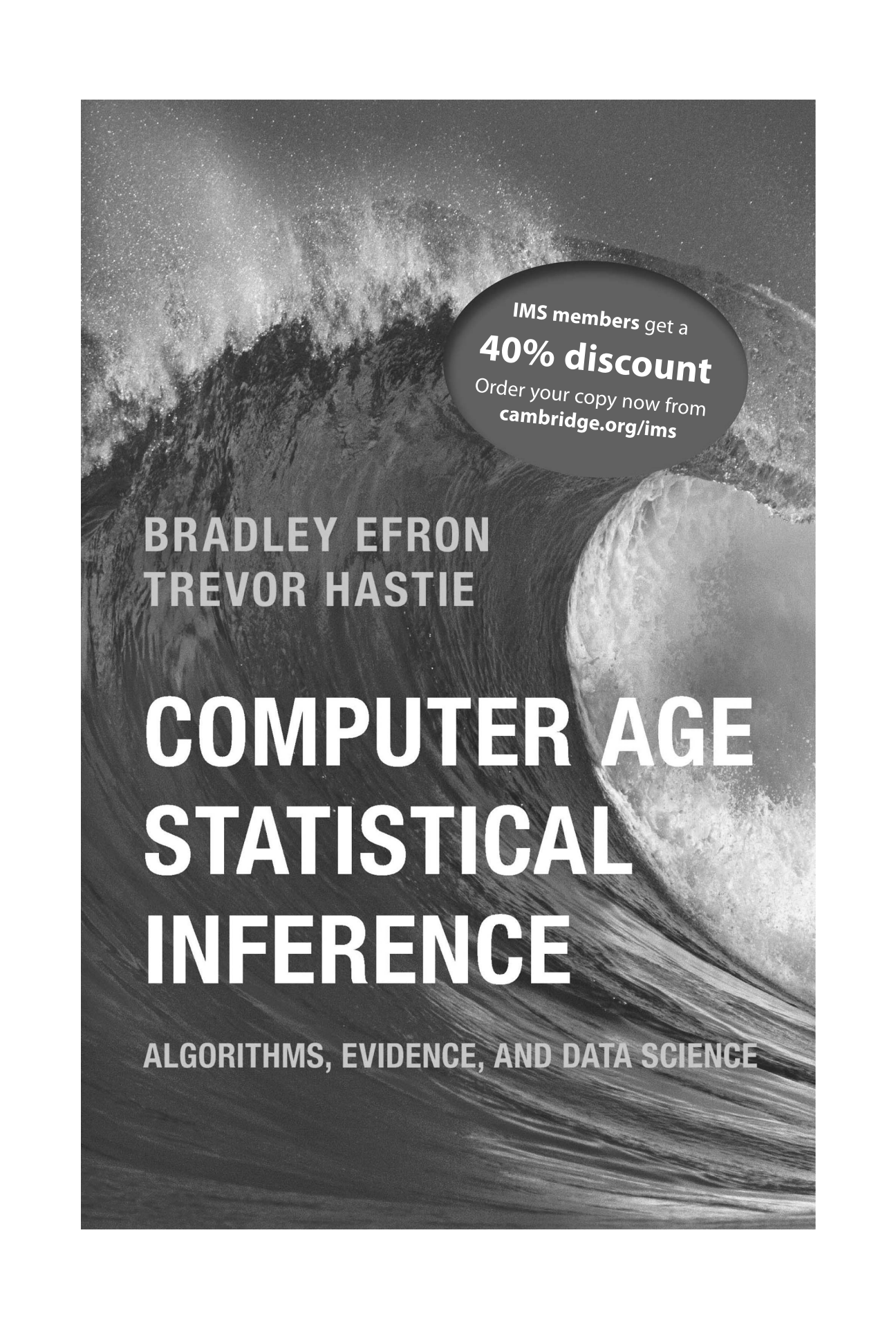
REFERENCES

- ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. [MR2549554](#) <https://doi.org/10.1214/09-AOS689>
- ANDREWS, S. (2011). In-close2, a high performance formal concept miner. In *International Conference on Conceptual Structures* 50–62. Springer, Berlin.
- ASGARIAN, N. and GREINER, R. (2007). Using rank-one biclusters to classify microarray data. Technical Report CPDC-TR-9906-010, Univ. Alberta.
- BEN-DOR, A., CHOR, B., KARP, R. and YAKHINI, Z. (2003). Discovering local structure in gene expression data: The order-preserving submatrix problem. *J. Comput. Biol.* **10** 373–384.
- BERGMANN, S., IHMELS, J. and BARKAI, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E* **67** 031902.
- BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.
- CALDAS, J. and KASKI, S. (2008). Bayesian biclustering with the plaid model. In *2008 IEEE Workshop on Machine Learning for Signal Processing* 291–296. IEEE, New York.
- CHENG, Y. and CHURCH, G. M. (2000). Biclustering of expression data. In *Ismb* **8** 93–103.
- CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90** 1313–1321. [MR1379473](#)
- CHU, L.-F., LENG, N., ZHANG, J., HOU, Z., MAMOTT, D., VEREIDE, D. T., CHOI, J., KENDZIORSKI, C., STEWART, R. et al. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* **17** 173.
- DE FRANÇA, F. O. (2016). A hash-based co-clustering algorithm for categorical data. *Expert Syst. Appl.* **64** 24–35.
- 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](#)
- DOMENICONI, C., PAPADOPOULOS, D., GUNOPULOS, D. and MA, S. (2004). Subspace clustering of high dimensional data. In *Proceedings of the Fourth SIAM International Conference on Data Mining* 517–521. SIAM, Philadelphia, PA. [MR2388480](#)

Key words and phrases. Bi-clustering, clustering, variable selection, categorical data, model selection, high dimensionality, genetics.

- GALIMBERTI, G. and SOFFRITTI, G. (2007). Model-based methods to identify multiple cluster structures in a data set. *Comput. Statist. Data Anal.* **52** 520–536. [MR2409999 https://doi.org/10.1016/j.csda.2007.02.019](https://doi.org/10.1016/j.csda.2007.02.019)
- GAN, G. and WU, J. (2004). Subspace clustering for high dimensional categorical data. *ACM SIGKDD Explor. Newsl.* **6** 87–94.
- GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. [MR1380810 https://doi.org/10.1093/biomet/82.4.711](https://doi.org/10.1093/biomet/82.4.711)
- GU, J. and LIU, J. S. (2008). Bayesian biclustering of gene expression data. *BMC Genomics* **9** S4.
- GUO, L. (2013). Bayesian Biclustering on Discrete Data: Variable Selection Methods. Ph.D. thesis, Dept. Statistics, Harvard Univ.
- HARTIGAN, J. A. (1972). Direct clustering of a data matrix. *J. Amer. Statist. Assoc.* **67** 123–129.
- HARTIGAN, J. A. (1975). *Clustering Algorithms. Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. [MR0405726](https://doi.org/10.1002/9780471914645)
- HOCHREITER, S., BODENHOFER, U., HEUSEL, M., MAYR, A., MITTERECKER, A., KASIM, A., KHAMI-AKOVA, T., VAN SANDEN, S., LIN, D. et al. (2010). FABIA: Factor analysis for bicluster acquisition. *Bioinformatics* **26** 1520–1527.
- HOFF, P. D. (2005). Subset clustering of binary sequences, with an application to genomic abnormality data. *Biometrics* **61** 1027–1036. [MR2216196 https://doi.org/10.1111/j.1541-0420.2005.00381.x](https://doi.org/10.1111/j.1541-0420.2005.00381.x)
- HOFF, P. D. (2006). Model-based subspace clustering. *Bayesian Anal.* **1** 321–344. [MR2221267 https://doi.org/10.1214/06-BA111](https://doi.org/10.1214/06-BA111)
- HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.
- JACKSON, J. E. (2005). *A User's Guide to Principal Components* **587**. Wiley, New York.
- JOHNSON, S. C. (1967). Hierarchical clustering schemes. *Psychometrika* **32** 241–254.
- KERIBIN, C., BRAULT, V., CELEUX, G. and GOVAERT, G. (2015). Estimation and selection for the latent block model on categorical data. *Stat. Comput.* **25** 1201–1216. [MR3401881 https://doi.org/10.1007/s11222-014-9472-2](https://doi.org/10.1007/s11222-014-9472-2)
- KIM, S., TADESSE, M. G. and VANNUCCI, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93** 877–893. [MR2285077 https://doi.org/10.1093/biomet/93.4.877](https://doi.org/10.1093/biomet/93.4.877)
- KISELEV, V. Y., KIRSCHNER, K., SCHAUB, M. T., ANDREWS, T., YIU, A., CHANDRA, T., NATARAJAN, K. N., REIK, W., BARAHONA, M. et al. (2017). SC3: Consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14** 483–486.
- LABIOD, L. and NADIF, M. (2011). Co-clustering for binary and categorical data with maximum modularity. In *2011 IEEE 11th International Conference on Data Mining* 1140–1145. IEEE, New York.
- LAZZERONI, L. and OWEN, A. (2002). Plaid models for gene expression data. *Statist. Sinica* **61**–86.
- LEE, M., SHEN, H., HUANG, J. Z. and MARRON, J. S. (2010). Biclustering via sparse singular value decomposition. *Biometrics* **66** 1087–1095. [MR2758496 https://doi.org/10.1111/j.1541-0420.2010.01392.x](https://doi.org/10.1111/j.1541-0420.2010.01392.x)
- LI, Y., JOURDAIN, A. A., CALVO, S. E., LIU, J. S. and MOOTHA, V. K. (2017). CLIC, a tool for expanding biological pathways based on co-expression across thousands of datasets. *PLoS Comput. Biol.* **13** e1005653.
- LI, Z., CHANG, C., KUNDU, S. and LONG, Q. (2020). Bayesian generalized biclustering analysis via adaptive structured shrinkage. *Biostatistics* **21** 610–624. [MR4120343 https://doi.org/10.1093/biostatistics/kxy081](https://doi.org/10.1093/biostatistics/kxy081)
- LIU, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Amer. Statist. Assoc.* **89** 958–966. [MR1294740](https://doi.org/10.1080/01621459.1994.1047740)
- LIU, J. S., ZHANG, J. L., PALUMBO, M. J. and LAWRENCE, C. E. (2003). Bayesian clustering with variable and transformation selections. *Bayesian Statistics* **7** 249–275.
- MACA-MEYER, N., GONZÁLEZ, A. M., LARRUGA, J. M., FLORES, C. and CABRERA, V. M. (2001). Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet.* **2** 1–8.
- MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)* 281–297. Univ. California Press, Berkeley, CA. [MR0214227](https://doi.org/10.1007/978-1-4613-9626-3_16)
- MARBAC, M. and SEDKI, M. (2017a). Variable selection for mixed data clustering: A model-based approach. ArXiv preprint. Available at [arXiv:1703.02293](https://arxiv.org/abs/1703.02293).
- MARBAC, M. and SEDKI, M. (2017b). Variable selection for model-based clustering using the integrated complete-data likelihood. *Stat. Comput.* **27** 1049–1063. [MR3627562 https://doi.org/10.1007/s11222-016-9670-1](https://doi.org/10.1007/s11222-016-9670-1)
- MARBAC, M. and VANDEWALLE, V. (2019). A tractable multi-partitions clustering. *Comput. Statist. Data Anal.* **132** 167–179. [MR3913142 https://doi.org/10.1016/j.csda.2018.06.013](https://doi.org/10.1016/j.csda.2018.06.013)
- NARASIMHAN, V. M., PATTERSON, N., MOORJANI, P., ROHLAND, N., BERNARDOS, R., MALICK, S., LAZARIDIS, I., NAKATSUKA, N., OLALDE, I. et al. (2019). The formation of human populations in South and central Asia. *Science* **365**.

- NEUWALD, A. F., KANNAN, N., POLEKSIC, A., HATA, N. and LIU, J. S. (2003). Ran's C-terminal, basic patch, and nucleotide exchange mechanisms in light of a canonical structure for Rab, Rho, Ras, and Ran GTPases. *Genome Res.* **13** 673–692.
- PADILHA, V. A. and CAMPELLO, R. J. (2017). A systematic comparative evaluation of biclustering techniques. *BMC Bioinform.* **18** 1–25.
- PENSA, R. G., ROBARDET, C. and BOULICAUT, J.-F. (2005). A bi-clustering framework for categorical data. In *European Conference on Principles of Data Mining and Knowledge Discovery* 643–650. Springer, Berlin.
- POSTH, C., RENAUD, G., MITTNIK, A., DRUCKER, D. G., ROUGIER, H., CUPILLARD, C., VALENTIN, F., THEVENET, C., FURTWÄNGLER, A. et al. (2016). Pleistocene mitochondrial genomes suggest a single major dispersal of non-africans and a late glacial population turnover in Europe. *Curr. Biol.* **26** 827–833.
- PRITCHARD, J. K., STEPHENS, M. and DONNELLY, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155** 945–959.
- PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81** 559–575.
- RAFTERY, A. E. and DEAN, N. (2006). Variable selection for model-based clustering. *J. Amer. Statist. Assoc.* **101** 168–178. [MR2268036 https://doi.org/10.1198/016214506000000113](https://doi.org/10.1198/016214506000000113)
- RAGHAVAN, M., SKOGLUND, P., GRAF, K. E., METSPALU, M., ALBRECHTSEN, A., MOLTKE, I., RASMUSSEN, S., STAFFORD JR., T. W., ORLANDO, L. et al. (2014). Upper palaeolithic Siberian genome reveals dual ancestry of native Americans. *Nature* **505** 87–91.
- REICH, D., THANGARAJ, K., PATTERSON, N., PRICE, A. L. and SINGH, L. (2009). Reconstructing Indian population history. *Nature* **461** 489–494.
- SANS, M. (2000). Admixture studies in Latin America: From the 20th to the 21st century. *Human Biology* 155–177.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. [MR0468014](https://doi.org/10.1214/aos/1176344948)
- SOKAL, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.* **38** 1409–1438.
- STUART, T., BUTLER, A., HOFFMAN, P., HAFEMEISTER, C., PAPALEXI, E., MAUCK III, W. M., HAO, Y., STOECKIUS, M., SMIBERT, P. et al. (2019). Comprehensive integration of single-cell data. *Cell* **177** 1888–1902.
- TADESSE, M. G., SHA, N. and VANNUCCI, M. (2005). Bayesian variable selection in clustering high-dimensional data. *J. Amer. Statist. Assoc.* **100** 602–617. [MR2160563 https://doi.org/10.1198/016214504000001565](https://doi.org/10.1198/016214504000001565)
- TANAY, A., SHARAN, R. and SHAMIR, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics* **18** S136–S144.
- WU, J. (2017). Bayesian inference of relationships. Ph.D. Thesis, Dept. Statistics, Harvard Univ.
- YAN, H., WU, J., LI, Y. and LIU, J. S. (2022). Supplement to “Bayesian bi-clustering methods with applications in computational biology.” <https://doi.org/10.1214/22-AOAS1622SUPP>
- YANG, W.-H., DAI, D.-Q. and YAN, H. (2007). Biclustering of microarray data based on singular value decomposition. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* 194–205. Springer, Berlin.



IMS members get a
40% discount

Order your copy now from
cambridge.org/ims

BRADLEY EFRON
TREVOR HASTIE

COMPUTER AGE STATISTICAL INFERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE