

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

Discovering interactions using covariate informed random partition models GARRITT L. PAGE, FERNANDO A. QUINTANA AND GARY L. ROSNER	1
Late 19th century navigational uncertainties and their influence on sea surface temperature estimates.....CHENGUANG DAI, DUO CHAN, PETER HUYBERS AND NATESH PILLAI	22
Analyzing second order stochasticity of neural spiking under stimuli-bundle exposure CHRIS GLYNN, SURYA T. TOKDAR, AZEEM ZAMAN, VALERIA C. CARUSO, JEFF T. MOHL, SHAWN M. WILLETT AND JENNIFER M. GROH	41
Integrative network learning for multimodality biomarker data SHANGHONG XIE, DONGLIN ZENG AND YUANJIA WANG	64
Point process modeling of drug overdoses with heterogeneous and missing data XUEYING LIU, JEREMY CARTER, BRAD RAY AND GEORGE MOHLER	88
An approximate best prediction approach to small area estimation for sheet and rill erosion under informative sampling EMILY BERG AND JAE-KWANG KIM	102
Robust inference when combining inverse-probability weighting and multiple imputation to address missing data with application to an electronic health records-based study of bariatric surgery TANAYOTT THAWEETHAI, DAVID E. ARTERBURN, KAREN J. COLEMAN AND SEBASTIEN HANEUSE	126
Spike-and-slab Lasso biclustering GEMMA E. MORAN, VERONIKA ROČKOVÁ AND EDWARD I. GEORGE	148
Random-effects meta-analysis of Phase I dose-finding studies using stochastic process priors .. MORENO URSINO, CHRISTIAN RÖVER, SARAH ZOHAR AND TIM FRIEDE	174
Model free estimation of graphical model using gene expression data JENNY YANG, YANG LIU, YUFENG LIU AND WEI SUN	194
An evaluation of statistical methods for aggregate patterns of replication failure JACOB M. SCHAUER, KAITLYN G. FITZGERALD, SARAH PEKO-SPICER, MENA C. R. WHALEN, RITA ZEJNULLAHI AND LARRY V. HEDGES	208
Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models ZHENYU ZHANG, AKIHIKO NISHIMURA, PAUL BASTIDE, XIANG JI, REBECCA P. PAYNE, PHILIP GOULDER, PHILIPPE LEMEY AND MARC A. SUCHARD	230
Monitoring vaccine safety by studying temporal variation of adverse events using vaccine adverse event reporting system JING HUANG, YI CAI, JINGCHENG DU, RUOSHA LI, SUSAN S. ELLENBERG, SEAN HENNESSY, CUI TAO AND YONG CHEN	252
The role of scale in the estimation of cell-type proportions GREGORY J. HUNT AND JOHANN A. GAGNON-BARTSCH	270
Two-phase sampling strategies for design-based mapping of continuous spatial populations in environmental surveys .. LORENZO FATTORINI, SARA FRANCESCHI, MARZIA MARCHESELLI, CATERINA PISANI AND LUCA PRATELLI	287

continued

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—*Continued from front cover*

- A regression discontinuity design for ordinal running variables: Evaluating central bank purchases of corporate bonds FAN LI, ANDREA MERCATANTI, TANELI MÄKINEN AND ANDREA SILVESTRINI 304
- Spatial distributed lag data fusion for estimating ambient air pollution
JOSHUA L. WARREN, MARIE LYNN MIRANDA, JOSHUA L. TOOTOO, CLAIRE E. OSGOOD AND MICHELLE L. BELL 323
- Estimation of cell lineage trees by maximum-likelihood phylogenetics JEAN FENG, WILLIAM S. DEWITT III, AARON MCKENNA, NOAH SIMON, AMY D. WILLIS AND FREDERICK A. MATSEN IV 343
- BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic QINGYUAN ZHAO, NIANQIAO JU, SERGIO BACALLADO AND RAJEN D. SHAH 363
- BayICE: A Bayesian hierarchical model for semireference-based deconvolution of bulk transcriptomic data AN-SHUN TAI, GEORGE C. TSENG AND WEN-PING HSIEH 391
- A multiple imputation procedure for record linkage and causal inference to estimate the effects of home-delivered meals MINGYANG SHAN, KALI S. THOMAS AND ROEE GUTMAN 412
- Accounting for smoking in forecasting mortality and life expectancy
YICHENG LI AND ADRIAN E. RAFTERY 437
- High-fidelity hurricane surge forecasting using emulation and sequential experiments
MATTHEW PLUMLEE, TAYLOR G. ASHER, WON CHANG AND MATTHEW V. BILSKIE 460
- Model-based feature selection and clustering of RNA-seq data for unsupervised subtype discovery DAVID K. LIM, NAIM U. RASHID AND JOSEPH G. IBRAHIM 481
- An empirical Bayes change-point model for transcriptome time-course data
TIAN TIAN, RUIHUA CHENG AND ZHI WEI 509

THE ANNALS OF APPLIED STATISTICS

Vol. 15, No. 1, pp. 1–526 March 2021

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

IMS OFFICERS

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

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

Past President: Susan Murphy, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

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

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

Program Secretary: Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027-5927, USA

IMS PUBLICATIONS

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

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

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

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

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

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

The Annals of Applied Statistics [ISSN 1932-6157 (print); ISSN 1941-7330 (online)], Volume 15, Number 1, March 2021. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage pending at Cleveland, Ohio, and at additional mailing offices.

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

DISCOVERING INTERACTIONS USING COVARIATE INFORMED RANDOM PARTITION MODELS

BY GARRITT L. PAGE¹, FERNANDO A. QUINTANA² AND GARY L. ROSNER³

¹Department of Statistics, Brigham Young University, page@stat.byu.edu

²Departamento de Estadística, Pontificia Universidad Católica de Chile, quintana@mat.uc.cl

³Department of Biostatistics, Johns Hopkins University, grosner1@jhmi.edu

Combination chemotherapy treatment regimens created for patients diagnosed with childhood acute lymphoblastic leukemia have had great success in improving cure rates. Unfortunately, patients prescribed these types of treatment regimens have displayed susceptibility to the onset of osteonecrosis. Some have suggested that this is due to pharmacokinetic interaction between two agents in the treatment regimen (asparaginase and dexamethasone) and other physiological variables. Determining which physiological variables to consider when searching for interactions in scenarios like these, minus a priori guidance, has proved to be a challenging problem, particularly if interactions influence the response distribution in ways beyond shifts in expectation or dispersion only. In this paper we propose an exploratory technique that is able to discover associations between covariates and responses in a general way. The procedure connects covariates to responses flexibly through dependent random partition distributions and then employs machine learning techniques to highlight potential associations found in each cluster. We provide a simulation study to show utility and apply the method to data produced from a study dedicated to learning which physiological predictors influence severity of osteonecrosis multiplicatively.

REFERENCES

- AGRAWAL, R., MANNILA, H., SRIKANT, R., TOIVONEN, H. and VERKAMO, A. I. (1996). Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining* (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, eds.) 307–328. American Association for Artificial Intelligence, Menlo Park, CA, USA.
- AGRAWAL, R., TRIPPE, B., HUGGINS, J. and BRODERICK, T. (2019). The kernel interaction trick: Fast Bayesian discovery of pairwise interactions in high dimensions. In *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.). *Proceedings of Machine Learning Research* **97** 141–150. PMLR, Long Beach, CA.
- AMERICAN CANCER SOCIETY (2018). Survival Rates for Childhood Leukemias. <https://www.cancer.org/cancer/leukemia-in-children/detection-diagnosis-staging/survival-rates.html> [Accessed: 18 March 2018].
- BAO, J. and HANSON, T. E. (2015). Bayesian nonparametric multivariate ordinal regression. *Canad. J. Statist.* **43** 337–357. MR3388321 <https://doi.org/10.1002/cjs.11253>
- BARCELLA, W., DE IORIO, M., FAVARO, S. and ROSNER, G. L. (2018). Dependent generalized Dirichlet process priors for the analysis of acute lymphoblastic leukemia. *Biostatistics* **19** 342–358. MR3984987 <https://doi.org/10.1093/biostatistics/kxx042>
- BARRERA-GÓMEZ, J., AGIER, L., PORTENGEN, L., CHADEAU-HYAM, M., GIORGIS-ALLEMAND, L., SIROUX, V., ROBINSON, O., VLAANDEREN, J., GONZÁLEZ, J. R. et al. (2017). A systematic comparison of statistical methods to detect interactions in exposome-health associations. *Environ. Health* **16** 74.
- BERGER, J. O., WANG, X. and SHEN, L. (2014). A Bayesian approach to subgroup identification. *J. Biopharm. Statist.* **24** 110–129. MR3196130 <https://doi.org/10.1080/10543406.2013.856026>
- BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A LASSO for hierarchical interactions. *Ann. Statist.* **41** 1111–1141. MR3113805 <https://doi.org/10.1214/13-AOS1096>

- CHEN, T. and GUESTIN, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16 785–794. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/2939672.2939785>
- CHEN, Y. and HANSON, T. E. (2014). Bayesian nonparametric k -sample tests for censored and uncensored data. *Comput. Statist. Data Anal.* **71** 335–346. [MR3131974 https://doi.org/10.1016/j.csda.2012.11.003](https://doi.org/10.1016/j.csda.2012.11.003)
- CHEN, T., HE, T., BENESTY, M., KHOTILOVICH, V., TANG, Y., CHO, H., CHEN, K., MITCHELL, R., CANO, I. et al. (2019). xgboost: Extreme gradient boosting. R package version 0.90.0.2.
- CHUNG, Y. and DUNSON, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *J. Amer. Statist. Assoc.* **104** 1646–1660. [MR2750582 https://doi.org/10.1198/jasa.2009.tm08302](https://doi.org/10.1198/jasa.2009.tm08302)
- DAHL, D. B. (2020). salso: Sequentially-allocated latent structure optimization. R package version 0.1.11.
- DE IORIO, M., MÜLLER, P., ROSNER, G. L. and MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99** 205–215. [MR2054299 https://doi.org/10.1198/016214504000000205](https://doi.org/10.1198/016214504000000205)
- DU, J. and LINERO, A. R. (2018). Interaction detection with Bayesian decision tree ensembles. [arXiv:1809.08524](https://arxiv.org/abs/1809.08524).
- DUNSON, D. B., PILLAI, N. and PARK, J.-H. (2007). Bayesian density regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 163–183. [MR2325270 https://doi.org/10.1111/j.1467-9868.2007.00582.x](https://doi.org/10.1111/j.1467-9868.2007.00582.x)
- FAN, J., YAO, Q. and TONG, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83** 189–206. [MR1399164 https://doi.org/10.1093/biomet/83.1.189](https://doi.org/10.1093/biomet/83.1.189)
- FERRARI, F. and DUNSON, D. B. (2019). Bayesian factor analysis for inference on interactions. [arXiv:1904.11603v1](https://arxiv.org/abs/1904.11603v1).
- 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](https://doi.org/10.1111/rssa.12378)
- GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–374.
- HAHSLER, M., BUCHTA, C., GRUEN, B. and HORNIK, K. (2015). arules: Mining association rules and frequent itemsets. R package version 1.3-1.
- HAN, J., KAMBER, M. and PEI, J. (2012). *Data Mining Concepts and Techniques*, 1st ed. Elsevier.
- 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](https://doi.org/10.1007/978-0-387-84858-7)
- HEBA, I., AMANY, A., AHMED, S. E. and AMR, S. (2014). Novel data-mining methodologies for detecting drug-drug interactions: A review of pharmacovigilance literature. In *Advances in Environmental Sciences, Development and Chemistry* (W. L. Staff, ed.) 301–314. Wseas LLC.
- HENDERSON, N. C., LOUIS, T. A., ROSNER, G. L. and VARADHAN, R. (2020). Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *Biostatistics* **21** 50–68. [MR4043845 https://doi.org/10.1093/biostatistics/kxy028](https://doi.org/10.1093/biostatistics/kxy028)
- HU, J., JOSHI, A. and JOHNSON, V. E. (2009). Log-linear models for gene association. *J. Amer. Statist. Assoc.* **104** 597–607. [MR2751441 https://doi.org/10.1198/jasa.2009.0025](https://doi.org/10.1198/jasa.2009.0025)
- ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. [MR2163158 https://doi.org/10.1214/009053604000001147](https://doi.org/10.1214/009053604000001147)
- ISHWARAN, H., RAO, J. S. and KOGALUR, U. B. (2013). spikeslab: Prediction and variable selection using spike and slab regression. R package version 1.1.5.
- KAPELNER, A. and BLEICH, J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. *J. Stat. Softw.* **70** 4.
- KARBOWIAK, E. and BIECEK, P. (2019). EIX: Explain interactions in 'XGBoost'. R package version 1.0.
- KAWEDIA, J. D., KASTE, S. C., PEI, D., PANETTA, J. C., CAI, X., CHENG, C., NEALE, G., HOWARD, S. C., EVANS, W. E. et al. (2011). Pharmacokinetic, pharmacodynamic, and pharmacogenetic determinants of osteonecrosis in children with acute lymphoblastic leukemia. *Blood* **117** 2340–2347.
- KOTTAS, A., MÜLLER, P. and QUINTANA, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *J. Comput. Graph. Statist.* **14** 610–625. [MR2170204 https://doi.org/10.1198/106186005X63185](https://doi.org/10.1198/106186005X63185)
- LIM, M. and HASTIE, T. (2015). Learning interactions via hierarchical group-lasso regularization. *J. Comput. Graph. Statist.* **24** 627–654. [MR3397226 https://doi.org/10.1080/10618600.2014.938812](https://doi.org/10.1080/10618600.2014.938812)
- LIM, M. and HASTIE, T. (2019). glinternet: Learning interactions via hierarchical group-lasso regularization. R package version 1.0.10.
- LIU, J., SIVAGANESAN, S., LAUD, P. W. and MÜLLER, P. (2017). A Bayesian subgroup analysis using collections of ANOVA models. *Biom. J.* **59** 746–766. [MR3672694 https://doi.org/10.1002/bimj.201600064](https://doi.org/10.1002/bimj.201600064)
- MACEACHERN, S. N. (2000). Dependent Dirichlet processes. Technical Report, Ohio State Univ., Dept. Statistics.

- MITRA, R., MÜLLER, P. and JI, Y. (2017). Bayesian multiplicity control for multiple graphs. *Canad. J. Statist.* **45** 44–61. [MR3623463](#) <https://doi.org/10.1002/cjs.11305>
- MÜLLER, P., QUINTANA, F. and ROSNER, G. L. (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist.* **20** 260–278. Supplementary material available online. [MR2816548](#) <https://doi.org/10.1198/jcgs.2011.09066>
- MÜLLER, P., QUINTANA, F. A., JARA, A. and HANSON, T. (2015). *Bayesian Nonparametric Data Analysis. Springer Series in Statistics*. Springer, Cham. [MR3309338](#) <https://doi.org/10.1007/978-3-319-18968-0>
- 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>
- PAGE, G. L. and QUINTANA, F. A. (2018). Calibrating covariate informed product partition models. *Stat. Comput.* **28** 1009–1031. [MR3835631](#) <https://doi.org/10.1007/s11222-017-9777-z>
- PAGE, G. L., QUINTANA, F. A. and ROSNER, G. L. (2021a). Supplement to “Discovering interactions using covariate informed random partition models.” <https://doi.org/10.1214/20-AOAS1372SUPPA>
- PAGE, G. L., QUINTANA, F. A. and ROSNER, G. L. (2021b). Source code for “Discovering interactions using covariate informed random partition models.” <https://doi.org/10.1214/20-AOAS1372SUPPB>
- R CORE TEAM (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- REICH, B. J., KALENDRA, E., STORLIE, C. B., BONDELL, H. D. and FUENTES, M. (2012). Variable selection for high dimensional Bayesian density estimation: Application to human exposure simulation. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 47–66. [MR2877584](#) <https://doi.org/10.1111/j.1467-9876.2011.00772.x>
- SCHNELL, P. M., TANG, Q., OFFEN, W. W. and CARLIN, B. P. (2016). A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics* **72** 1026–1036. [MR3591587](#) <https://doi.org/10.1111/biom.12522>
- SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. [MR2722450](#) <https://doi.org/10.1214/10-AOS792>
- SHEN, W. and GHOSAL, S. (2016). Adaptive Bayesian density regression for high-dimensional data. *Bernoulli* **22** 396–420. [MR3449788](#) <https://doi.org/10.3150/14-BEJ663>
- SIMON, R. (2002). Bayesian subset analysis: Application to studying treatment-by-gender interactions. *Stat. Med.* **21** 2909–2916.
- SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75** 317–343.
- SU, X., PEÑA, A. T., LIU, L. and LEVINE, R. A. (2018). Random forests of interaction trees for estimating individualized treatment effects in randomized trials. *Stat. Med.* **37** 2547–2560. [MR3824516](#) <https://doi.org/10.1002/sim.7660>
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- TOKDAR, S. T., ZHU, Y. M. and GHOSH, J. K. (2010). Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Anal.* **5** 319–344. [MR2719655](#) <https://doi.org/10.1214/10-BA605>
- VARADHAN, R. and WANG, S.-J. (2014). Standardization for subgroup analysis in randomized controlled trials. *J. Biopharm. Statist.* **24** 154–167. [MR3196132](#) <https://doi.org/10.1080/10543406.2013.856023>
- VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*, Fourth ed. Springer, New York. ISBN 0-387-95457-0. [MR1337030](#) <https://doi.org/10.1007/978-1-4899-2819-1>
- WADE, S. and GHAHRAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Anal.* **13** 559–626. With discussion and a reply by the authors. [MR3807860](#) <https://doi.org/10.1214/17-BA1073>

LATE 19TH CENTURY NAVIGATIONAL UNCERTAINTIES AND THEIR INFLUENCE ON SEA SURFACE TEMPERATURE ESTIMATES

BY CHENGUANG DAI^{1,*}, DUO CHAN^{2,‡}, PETER HUYBERS^{2,§} AND NATESH PILLAI^{1,†}

¹Department of Statistics, Harvard University, *chenguangdai@g.harvard.edu; †pillai@fas.harvard.edu

²Department of Earth and Planetary Sciences, Harvard University, ‡duochan@g.harvard.edu; §phuybers@fas.harvard.edu

Accurate estimates of historical changes in sea surface temperatures (SSTs) and their uncertainties are important for documenting and understanding historical changes in climate. A source of uncertainty that has not previously been quantified in historical SST estimates stems from position errors. A Bayesian inference framework is proposed for quantifying errors in reported positions and their implications on SST estimates. The analysis framework is applied to data from the International Comprehensive Ocean-Atmosphere Data Set (ICOADS3.0) in 1885, a time when astronomical and chronometer estimation of position was common but predated the use of radio signals. Focus is upon a subset of 943 ship tracks from ICOADS3.0 that report their position every two hours to a precision of 0.01° longitude and latitude. These data are interpreted as positions determined by dead reckoning that are periodically updated by celestial correction techniques. The posterior medians of uncertainties in celestial correction are 33.1 km (0.30° on the equator) in longitude and 24.4 km (0.22°) in latitude, respectively. Celestial navigation uncertainties being smaller in latitude than longitude is qualitatively consistent with the relative difficulty of obtaining astronomical estimates. The posterior medians for two-hourly dead reckoning uncertainties are 19.2% for ship speed and 13.2° for ship heading, leading to random position uncertainties with median 0.18° (20 km on the equator) in longitude and 0.15° (17 km) in latitude. Reported ship tracks also contain systematic position uncertainties relating to precursor dead-reckoning positions not being updated after obtaining celestial position estimates, indicating that more accurate positions can be provided for SST observations. Finally, we translate position errors into SST uncertainties by sampling an ensemble of SSTs from the Multiscale Ultrahigh Resolution Sea Surface Temperature (MURSST) data set. Evolving technology for determining ship position, heterogeneous reporting and archiving of position information, and seasonal and spatial changes in navigational uncertainty and SST gradients together imply that accounting for positional error in SST estimates over the span of the instrumental record will require substantial additional effort.

REFERENCES

- BOWDITCH, J. I. (1906). *American Practical Navigator*. US Government Printing Office. Number 9.
- CARELLA, G., KENT, E. C. and BERRY, D. I. (2017). A probabilistic approach to ship voyage reconstruction in ICOADS. *Int. J. Climatol.* **37** 2233–2247.
- CERVONE, D. and PILLAI, N. S. (2015). Gaussian process regression with location errors. [arXiv:1506.08256](https://arxiv.org/abs/1506.08256).
- CHAN, D. and HUYBERS, P. (2019). Systematic differences in bucket sea surface temperature measurements amongst nations identified using a linear-mixed-effect method. *J. Climate*.
- CHAN, D. and WU, Q. (2015). Attributing observed SST trends and subcontinental land warming to anthropogenic forcing during 1979–2005. *J. Climate* **28** 3152–3170.
- CHAN, D., KENT, E. C., BERRY, D. I. and HUYBERS, P. (2019). Correcting datasets leads to more homogeneous early-twentieth-century sea surface warming. *Nature* **571** 393–397. <https://doi.org/10.1038/s41586-019-1349-2>

- CHIN, T. M., VAZQUEZ-CUERVO, J. and ARMSTRONG, E. M. (2017). A multi-scale high-resolution analysis of global sea surface temperature. *Remote Sens. Environ.* **200** 154–169.
- DEE, D. P., UPPALA, S. M., SIMMONS, A. J., BERRISFORD, P., POLI, P., KOBAYASHI, S., ANDRAE, U., BALMASEDA, M. A., BALSAMO, G. et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137** 553–597.
- DIRECTOR, H. and BORN, L. (2015). Connecting point-level and gridded moments in the analysis of climate data. *J. Climate* **28** 3496–3510.
- FOLLAND, C. (2005). Assessing bias corrections in historical sea surface temperature using a climate model. *Int. J. Climatol.* **25** 895–911.
- FREEMAN, E., WOODRUFF, S. D., WORLEY, S. J., LUBKER, S. J., KENT, E. C., ANGEL, W. E., BERRY, D. I., BROHAN, P., EASTMAN, R. et al. (2017). ICOADS Release 3.0: A major update to the historical marine climate record. *Int. J. Climatol.* **37** 2211–2232.
- FRIED, W. R. (1977). A comparative performance analysis of modern ground-based, air-based, and satellite-based radio navigation systems. *Navigation* **24** 48–58.
- GREGORY, J. M., STOUFFER, R. J., RAPER, S. C. B., STOTT, P. A. and RAYNER, N. A. (2002). An observationally based estimate of the climate sensitivity. *J. Climate* **15** 3117–3121.
- HUANG, B., THORNE, P. W., BANZON, V. F., BOYER, T., CHEPURIN, G., LAWRIE, J. H., MENNE, M. J., SMITH, T. M., VOSE, R. S. et al. (2017). Extended reconstructed sea surface temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *J. Climate* **30** 8179–8205.
- INGLEBY, B. (2010). Factors affecting ship and buoy data quality: A data assimilation perspective. *J. Atmos. Ocean. Technol.* **27** 1476–1489.
- KENNEDY, J. J. (2014). A review of uncertainty in in situ measurements and data sets of sea surface temperature. *Rev. Geophys.* **52** 1–32.
- KENNEDY, J. J., SMITH, R. O. and RAYNER, N. A. (2012). Using AATSR data to assess the quality of in situ sea-surface temperature observations for climate studies. *Remote Sens. Environ.* **116** 79–92.
- KENNEDY, J. J., RAYNER, N. A., SMITH, R. O., PARKER, D. E. and SAUNBY, M. (2011a). Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *J. Geophys. Res., Atmos.* **116**.
- KENNEDY, J. J., RAYNER, N. A., SMITH, R. O., PARKER, D. E. and SAUNBY, M. (2011b). Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *J. Geophys. Res., Atmos.* **116**.
- KENT, E. C. and CHALLENGER, P. G. (2006). Toward estimating climatic trends in SST. Part II: Random errors. *J. Atmos. Ocean. Technol.* **23** 476–486.
- KENT, E. C., CHALLENGER, P. G. and TAYLOR, P. K. (1999). A statistical determination of the random observational errors present in voluntary observing ships meteorological reports. *J. Atmos. Ocean. Technol.* **16** 905–914.
- KENT, E. C., KENNEDY, J. J., SMITH, T. M., HIRAHARA, S., HUANG, B., KAPLAN, A., PARKER, D. E., ATKINSON, C. P., BERRY, D. I. et al. (2017). A call for new approaches to quantifying biases in observations of sea surface temperature. *Bull. Am. Meteorol. Soc.* **98** 1601–1616.
- MORICE, C. P., KENNEDY, J. J., RAYNER, N. A. and JONES, P. D. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res., Atmos.* **117**.
- SOBEL, D. (2007). *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time*. Bloomsbury Publishing USA.
- STAN DEVELOPMENT TEAM (2019). RStan: The R interface to Stan. R package version 2.19.2.
- WOODRUFF, S. D., DIAZ, H. F., ELMS, J. D. and WORLEY, S. J. (1998). COADS Release 2 data and metadata enhancements for improvements of marine surface flux fields. *Phys. Chem. Earth* **23** 517–526.
- WUNSCH, C. (2004). Gulf Stream safe if wind blows and Earth turns. *Nature* **428** 601. <https://doi.org/10.1038/428601c>

ANALYZING SECOND ORDER STOCHASTICITY OF NEURAL SPIKING UNDER STIMULI-BUNDLE EXPOSURE

BY CHRIS GLYNN¹, SURYA T. TOKDAR², AZEEM ZAMAN³, VALERIA C. CARUSO⁴,
JEFF T. MOHL^{5,*}, SHAWN M. WILLETT⁶ AND JENNIFER M. GROH^{5,†}

¹Zillow Research, christophergl@zillowgroup.com

²Department of Statistical Science, Duke University, surya.tokdar@duke.edu

³Department of Statistics, Harvard University, azaman@g.harvard.edu

⁴Center for Human Growth and Development, University of Michigan, vcarus@umich.edu

⁵Department of Psychology and Neuroscience, Duke University, *jeffrey.mohl@duke.edu; †jmgroh@duke.edu

⁶Department of Neurobiology, Duke University, shawn.willett@duke.edu

Conventional analysis of neuroscience data involves computing average neural activity over a group of trials and/or a period of time. This approach may be particularly problematic when assessing the response patterns of neurons to more than one simultaneously presented stimulus. In such cases the brain must represent each individual component of the stimuli bundle, but trial-and-time-pooled averaging methods are fundamentally unequipped to address the means by which multiitem representation occurs. We introduce and investigate a novel statistical analysis framework that relates the firing pattern of a single cell, exposed to a stimuli bundle, to the ensemble of its firing patterns under each constituent stimulus. Existing statistical tools focus on what may be called “first order stochasticity” in trial-to-trial variation in the form of unstructured noise around a fixed firing rate curve associated with a given stimulus. Our analysis is based upon the theoretical premise that exposure to a stimuli bundle induces additional stochasticity in the cell’s response pattern in the form of a stochastically varying recombination of its single stimulus firing rate curves. We discuss challenges to statistical estimation of such “second order stochasticity” and address them with a novel dynamic admixture point process (DAPP) model. DAPP is a hierarchical point process model that decomposes second order stochasticity into a Gaussian stochastic process and a random vector of interpretable features and facilitates borrowing of information on the latter across repeated trials through latent clustering. We illustrate the utility and accuracy of the DAPP analysis with synthetic data simulation studies. We present real-world evidence of second order stochastic variation with an analysis of monkey inferior colliculus recordings under auditory stimuli.

REFERENCES

- ADLER, R. J. and TAYLOR, J. E. (2009). *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer, New York. [MR2319516](#)
- CARUSO, V. C., MOHL, J. T., GLYNN, C., LEE, J., WILLETT, S. M., ZAMAN, A., EBIHARA, A. F., ESTRADA, R., FREIWALD, W. A., TOKDAR, S. T. and GROH, J. M. (2018). Single neurons may encode simultaneous stimuli by switching between activity patterns. *Nat. Commun.* **9** 2715.
- ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. [MR1340510](#)
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. [MR0350949](#)
- FRIEDMAN, J. H. (1984). A variable span smoother. Technical report, Stanford Univ. CA Lab for Computational Statistics.

Key words and phrases. Spike train, multiple stimuli, dynamic admixture of Poisson processes, Gaussian process, Dirichlet process, Bayesian inference.

- GERSTEIN, G. L. and KIANG, N. Y.-S. (1960). An approach to the quantitative analysis of electrophysiological data from single neurons. *Biophys. J.* **1** 15.
- GLYNN, C., TOKDAR, S., ZAMAN, A., CARUSO, V., MOHL, J., WILLETT, S. and GROH, J. (2021). Supplement to “Analyzing second order stochasticity of neural spiking under stimuli-bundle exposure.” <https://doi.org/10.1214/20-AOAS1383SUPP>
- GRIFFIN, J. E. (2010). Default priors for density estimation with mixture models. *Bayesian Anal.* **5** 45–64. <https://doi.org/10.1214/10-BA502>
- KASS, R. E. and VENTURA, V. (2001). A spike-train probability model. *Neural Comput.* **13** 1713–1720.
- KASS, R. E. and VENTURA, V. (2006). Spike count correlation increases with length of time interval in the presence of trial-to-trial variation. *Neural Comput.* **18** 2583–2591.
- KASS, R. E., VENTURA, V. and BROWN, E. N. (2005). Statistical issues in the analysis of neuronal data. *J. Neurophysiol.* **94** 8–25.
- KASS, R. E., VENTURA, V. and CAI, C. (2003). Statistical smoothing of neuronal data. *Netw. Comput. Neural Syst.* **14** 5–16.
- NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. <https://doi.org/10.2307/1390653>
- 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. <https://doi.org/10.1080/01621459.2013.829001>
- RAMEZAN, R., MARRIOTT, P. and CHENOURI, S. (2016). Skellam process with resetting: A neural spike train model. *Stat. Med.* **35** 5717–5729. <https://doi.org/10.1002/sim.7127>
- RAO, V. and TEH, Y. W. (2011). Gaussian process modulated renewal processes. In *Proceedings of the 24th International Conference on Neural Information Processing Systems* 2474–2482. Curran Associates Inc., Red Hook.
- SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. <https://doi.org/10.1007/BF02429533>
- SHIMAZAKI, H. and SHINOMOTO, S. (2007). A method for selecting the bin size of a time histogram. *Neural Comput.* **19** 1503–1527. <https://doi.org/10.1162/neco.2007.19.6.1503>
- TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. <https://doi.org/10.1198/016214506000000302>
- TOKDAR, S. T. and MARTIN, R. G. (2019). Bayesian test of normality versus a Dirichlet process mixture alternative. *Sankhya, Ser. B*. To appear.
- TOKDAR, S., XI, P., KELLY, R. C. and KASS, R. E. (2010). Detection of bursts in extracellular spike trains using hidden semi-Markov point process models. *J. Comput. Neurosci.* **29** 203–212.
- TRUCCOLO, W., EDEN, U. T., FELLOWS, M. R., DONOGHUE, J. P. and BROWN, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J. Neurophysiol.* **93** 1074–1089.
- VENTURA, V., CAI, C. and KASS, R. E. (2005). Trial-to-trial variability and its effect on time-varying dependency between two neurons. *J. Neurophysiol.* **94** 2928–2939.
- WOLPERT, R. L. and ICKSTADT, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85** 251–267. <https://doi.org/10.1093/biomet/85.2.251>

INTEGRATIVE NETWORK LEARNING FOR MULTIMODALITY BIOMARKER DATA

BY SHANGHONG XIE^{1,*}, DONGLIN ZENG² AND YUANJIA WANG^{1,†}

¹Department of Biostatistics, Mailman School of Public Health, Columbia University, *sx2168@cumc.columbia.edu; [†yw2016@cumc.columbia.edu](mailto:yw2016@cumc.columbia.edu)

²Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, dzeng@email.unc.edu

The biomarker networks measured by different modalities of data (e.g., structural magnetic resonance imaging (sMRI), diffusion tensor imaging (DTI)) may share the same true underlying biological model. In this work we propose a nodewise biomarker graphical model to leverage the shared mechanism between multimodality data to provide a more reliable estimation of the target modality network and account for the heterogeneity in networks due to differences between subjects and networks of external modality. Latent variables are introduced to represent the shared unobserved biological network, and the information from the external modality is incorporated to model the distribution of the underlying biological network. We propose an efficient approximation to the posterior expectation of the latent variables that reduces computational cost by at least 50%. The performance of the proposed method is demonstrated by extensive simulation studies and an application to construct gray matter brain atrophy network of Huntington's disease by using sMRI data and DTI data. The identified network connections are more consistent with clinical literature and better improve prediction in follow-up clinical outcomes and separate subjects into clinically meaningful subgroups with different prognosis than alternative methods.

REFERENCES

- ALEXANDER-BLOCH, A., GIEDD, J. N. and BULLMORE, E. (2013). Imaging structural co-variance between human brain regions. *Nat. Rev., Neurosci.* **14** 322–336.
- ALEXANDER-BLOCH, A., RAZNAHAN, A., BULLMORE, E. and GIEDD, J. (2013). The convergence of maturational change and structural covariance in human cortical networks. *J. Neurosci.* **33** 2889–2899.
- BETZEL, R. F., AVENA-KOENIGSBERGER, A., GOÑI, J., HE, Y., DE REUS, M. A., GRIFFA, A., VÉRTES, P. E., MIŠIĆ, B., THIRAN, J.-P. et al. (2016). Generative models of the human connectome. *NeuroImage* **124** 1054–1064.
- BULLMORE, E. and SPORNS, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev., Neurosci.* **10** 186–198. <https://doi.org/10.1038/nrn2575>
- 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](https://doi.org/10.1093/biomet/asn034)
- COPPEN, E. M., VAN DER GROND, J., HAFKEMEIJER, A., ROMBOUTS, S. A. R. B. and ROOS, R. A. C. (2016). Early grey matter changes in structural covariance networks in Huntington's disease. *NeuroImage Clin.* **12** 806–814. <https://doi.org/10.1016/j.nicl.2016.10.009>
- 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)
- 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.
- EPSKAMP, S. and FRIED, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychol. Methods* **23** 617–634. <https://doi.org/10.1037/met0000167>

Key words and phrases. Graphical models, network analysis, multi-modality data, structural co-variance network, white matter connectivity, Huntington's disease.

- FORNITO, A., ZALESKY, A. and BULLMORE, E. (2016). *Fundamentals of Brain Network Analysis*. Academic Press, San Diego.
- FOYCEL, R. and DRTON, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems* 604–612.
- FOYCEL, R. and DRTON, M. (2011). Bayesian model choice and information criteria in sparse generalized linear models. Preprint. Available at [arXiv:1112.5635](https://arxiv.org/abs/1112.5635).
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.
- GONG, G., HE, Y., CHEN, Z. J. and EVANS, A. C. (2012). Convergence and divergence of thickness correlations with diffusion connections across the human cerebral cortex. *NeuroImage* **59** 1239–1248.
- HAO, B., SUN, W. W., LIU, Y. and CHENG, G. (2017). Simultaneous clustering and estimation of heterogeneous graphical models. *J. Mach. Learn. Res.* **18** Paper No. 217, 58. [MR3827105](https://arxiv.org/abs/1703.09849)
- HE, Y., CHEN, Z. J. and EVANS, A. C. (2007). Small-world anatomical networks in the human brain revealed by cortical thickness from MRI. *Cereb. Cortex* **17** 2407–2419.
- HINNE, M., AMBROGIONI, L., JANSSEN, R. J., HESKES, T. and VAN GERVEN, M. A. J. (2014). Structurally-informed Bayesian functional connectivity analysis. *NeuroImage* **86** 294–305. <https://doi.org/10.1016/j.neuroimage.2013.09.075>
- 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](https://doi.org/10.1214/ss/1009212673)
- JOHNSON, E. B., REES, E. M., LABUSCHAGNE, I., DURR, A., LEAVITT, B. R., ROOS, R. A., REILMANN, R., JOHNSON, H., HOBBS, N. Z. et al. (2015). The impact of occipital lobe cortical thickness on cognitive task performance: An investigation in Huntington’s disease. *Neuropsychologia* **79** 138–146.
- KLÖPPEL, S., GREGORY, S., SCHELLER, E., MINKOVA, L., RAZI, A., DURR, A., ROOS, R. A., LEAVITT, B. R., PAPOUTSI, M. et al. (2015). Compensation in preclinical Huntington’s disease: Evidence from the track-on HD study. *EBioMedicine* **2** 1420–1429.
- LERCH, J. P., WORSLEY, K., SHAW, W. P., GREENSTEIN, D. K., LENROOT, R. K., GIEDD, J. and EVANS, A. C. (2006). Mapping anatomical correlations across cerebral cortex (MACACC) using cortical thickness from MRI. *NeuroImage* **31** 993–1003.
- MCCOLGAN, P., SEUNARINE, K. K., GREGORY, S., RAZI, A., PAPOUTSI, M., LONG, J. D., MILLS, J. A., JOHNSON, E., DURR, A. et al. (2017a). Topological length of white matter connections predicts their rate of atrophy in premanifest Huntington’s disease. *JCI Insight* **2**.
- MCCOLGAN, P., RAZI, A., GREGORY, S., SEUNARINE, K. K., DURR, A., ROOS, R. A. C., LEAVITT, B. R., SCAHILL, R. I., CLARK, C. A. et al. (2017b). Structural and functional brain network correlates of depressive symptoms in premanifest Huntington’s disease. *Hum. Brain Mapp.* **38** 2819–2829.
- NG, B., VAROQUAUX, G., POLINE, J.-B. and THIRION, B. (2012). A novel sparse graphical approach for multi-modal brain connectivity inference. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 707–714.
- PAULSEN, J. S., HAYDEN, M., STOUT, J. C., LANGBEHN, D. R., AYLWARD, E., ROSS, C. A., GUTTMAN, M., NANCE, M., KIEBURTZ, K. et al. (2006). Preparing for preventive clinical trials: The PREDICT-HD study. *Arch. Neurol.* **63** 883–890.
- PAULSEN, J., LANGBEHN, D., STOUT, J., AYLWARD, E., ROSS, C., NANCE, M., GUTTMAN, M., JOHNSON, S., MACDONALD, M. et al. (2008). Detection of Huntington’s disease decades before diagnosis: The predict-HD study. *J. Neurol. Neurosurg. Psychiatry* **79** 874–880.
- PINEDA-PARDO, J. A., BRUÑA, R., WOOLRICH, M., MARCOS, A., NOBRE, A. C., MAESTÚ, F. and VIDAU-RE, D. (2014). Guiding functional connectivity estimation by structural connectivity in MEG: An application to discrimination of conditions of mild cognitive impairment. *NeuroImage* **101** 765–777.
- RIZVI, B., NARKHEDE, A., LAST, B. S., BUDGE, M., TOSTO, G., MANLY, J. J., SCHUPF, N., MAYEUX, R. and BRICKMAN, A. M. (2018). The effect of white matter hyperintensities on cognition is mediated by cortical atrophy. *Neurobiol. Aging* **64** 25–32. <https://doi.org/10.1016/j.neurobiolaging.2017.12.006>
- ROSAS, H., HEVELONE, N., ZALETA, A., GREVE, D., SALAT, D. and FISCHL, B. (2005). Regional cortical thinning in preclinical Huntington disease and its relationship to cognition. *Neurology* **65** 745–747.
- ROSAS, H. D., SALAT, D. H., LEE, S. Y., ZALETA, A. K., PAPPU, V., FISCHL, B., GREVE, D., HEVELONE, N. and HERSCH, S. M. (2008). Cerebral cortex and the clinical expression of Huntington’s disease: Complexity and heterogeneity. *Brain* **131** 1057–1068.
- SCHMITT, J., LENROOT, R., WALLACE, G., ORDAZ, S., TAYLOR, K., KABANI, N., GREENSTEIN, D., LERCH, J., KENDLER, K. et al. (2008). Identification of genetically mediated cortical networks: A multi-variate study of pediatric twins and siblings. *Cereb. Cortex* **18** 1737–1747.
- SEELEY, W. W., CRAWFORD, R. K., ZHOU, J., MILLER, B. L. and GREICIUS, M. D. (2009). Neurodegenerative diseases target large-scale human brain networks. *Neuron* **62** 42–52.

- SPIEGELHALTER, D. J., ABRAMS, K. R. and MYLES, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation* **13**. Wiley, New York.
- STÄDLER, N., DONDELINGER, F., HILL, S. M., AKBANI, R., LU, Y., MILLS, G. B. and MUKHERJEE, S. (2017). Molecular heterogeneity at the network level: High-dimensional testing, clustering and a TCGA case study. *Bioinformatics* **33** 2890–2896. <https://doi.org/10.1093/bioinformatics/btx322>
- TABRIZI, S. J., LANGBEHN, D. R., LEAVITT, B. R., ROOS, R. A., DURR, A., CRAUFURD, D., KENNARD, C., HICKS, S. L., FOX, N. C. et al. (2009). Biological and clinical manifestations of Huntington’s disease in the longitudinal TRACK-HD study: Cross-sectional analysis of baseline data. *Lancet Neurol.* **8** 791–801.
- VÉRTES, P. E., ALEXANDER-BLOCH, A. F., GOGTAY, N., GIEDD, J. N., RAPOPORT, J. L. and BULLMORE, E. T. (2012). Simple models of human brain functional networks. *Proc. Natl. Acad. Sci. USA* **109** 5868–5873.
- WANG, Z., GU, Q., NING, Y. and LIU, H. (2015). High dimensional EM algorithm: Statistical optimization and asymptotic normality. In *Advances in Neural Information Processing Systems* 2521–2529.
- XIE, S., ZENG, D. and WANG, Y. (2021). Supplement to “Integrative network learning for multimodality biomarker data.” <https://doi.org/10.1214/20-AOAS1382SUPPA>, <https://doi.org/10.1214/20-AOAS1382SUPPB>
- ZHANG, Y., LONG, J. D., MILLS, J. A., WARNER, J. H., LU, W. and PAULSEN, J. S. (2011). Indexing disease progression at study entry with individuals at-risk for Huntington disease. *Am. J. Med. Genet., Part B Neuropsychiatr. Genet.* **156** 751–763.

POINT PROCESS MODELING OF DRUG OVERDOSES WITH HETEROGENEOUS AND MISSING DATA

BY XUEYING LIU^{1,*}, JEREMY CARTER^{2,‡}, BRAD RAY^{2,§} AND GEORGE MOHLER^{1,†}

¹Department of Computer and Information Science, Indiana University–Purdue University Indianapolis, *xl17@iu.edu; †gmohler@iupui.edu

²O'Neill School of Public and Environmental Affairs, Indiana University–Purdue University Indianapolis, ‡carterjg@iupui.edu; §bradray@iupui.edu

Opioid overdose rates have increased in the United States over the past decade and reflect a major public health crisis. Modeling and prediction of drug and opioid hotspots, where a high percentage of events fall in a small percentage of space–time, could help better focus limited social and health services. In this work we present a spatial-temporal point process model for drug overdose clustering. The data input into the model comes from two heterogeneous sources: (1) high volume emergency medical calls for service (EMS) records containing location and time but no information on the type of nonfatal overdose, and (2) fatal overdose toxicology reports from the coroner containing location and high-dimensional information from the toxicology screen on the drugs present at the time of death. We first use nonnegative matrix factorization to cluster toxicology reports into drug overdose categories, and we then develop an EM algorithm for integrating the two heterogeneous data sets, where the mark corresponding to overdose category is inferred for the EMS data and the high volume EMS data is used to more accurately predict drug overdose death hotspots. We apply the algorithm to drug overdose data from Indianapolis, showing that the point process defined on the integrated data out-performs point processes that use only coroner data (AUC improvement 0.81 to 0.85). We also investigate the extent to which overdoses are contagious, as a function of the type of overdose, while controlling for exogenous fluctuations in the background rate that might also contribute to clustering. We find that drug and opioid overdose deaths exhibit significant excitation with branching ratio ranging from 0.72 to 0.98.

REFERENCES

- ALEXANDER, M. J., KIANG, M. V. and BARBIERI, M. (2018). Trends in black and white opioid mortality in the United States, 1979–2015. *Epidemiology* **29** 707.
- CARTER, J. G., MOHLER, G. and RAY, B. (2018). Spatial concentration of opioid overdose deaths in Indianapolis: An application of the law of crime concentration at place to a public health epidemic. *J. Contemp. Crim. Justice*. <https://doi.org/10.1177/1043986218803527>
- CICCARONE, D. (2017). Fentanyl in the US heroin supply: A rapidly changing risk environment. *Int. J. Drug Policy* **46** 107–111.
- CICERO, T. J., ELLIS, M. S., SURRATT, H. L. and KURTZ, S. P. (2014). The changing face of heroin use in the United States: A retrospective analysis of the past 50 years. *JAMA Psychiatr.* **71** 821–826.
- FISHER, R., O'DONNELL, D., RAY, B. and RUSYNAK, D. (2016). Police officers can safely and effectively administer intranasal naloxone. *Prehosp. Emerg. Care* **20** 675–680. <https://doi.org/10.1080/10903127.2016.1182605>
- GLADDEN, R. M., MARTINEZ, P. and SETH, P. (2016). Fentanyl law enforcement submissions and increases in synthetic opioid-involved overdose deaths—27 states, 2013–2014. *Morb. Mort. Wkly. Rep.* **65** 837–843. <https://doi.org/10.15585/mmwr.mm6533a2>
- HEIDARYSAFA, M., KOWSARI, K., BARNES, L. and BROWN, D. (2018). Analysis of railway accidents' narratives using deep learning. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) 1446–1453. IEEE.

Key words and phrases. Point process, expectation maximization algorithm, semisupervised learning, nonnegative matrix factorization, opioid overdose.

- HIBDON, J. and GROFF, E. R. (2014). What you find depends on where you look: Using emergency medical services call data to target illicit drug use hot spots. *J. Contemp. Crim. Justice* **30** 169–185.
- HIBDON, J., TELEP, C. W. and GROFF, E. R. (2017). The concentration and stability of drug activity in Seattle, Washington using police and emergency medical services data. *J. Quant. Criminol.* **33** 497–517.
- JALAL, H., BUCHANICH, J. M., ROBERTS, M. S., BALMERT, L. C., ZHANG, K. and BURKE, D. S. (2018). Changing dynamics of the drug overdose epidemic in the United States from 1979 through 2016. *Science* **361** eaau1184.
- JONES, C. M., BALDWIN, G. T. and COMPTON, W. M. (2017). Recent increases in cocaine-related overdose deaths and the role of opioids. *Am. J. Publ. Health* **107** 430–432. <https://doi.org/10.2105/AJPH.2016.303627>
- JONES, C. M., EINSTEIN, E. B. and COMPTON, W. M. (2018). Changes in synthetic opioid involvement in drug overdose deaths in the United States, 2010–2016. *JAMA* **319** 1819–1821. <https://doi.org/10.1001/jama.2018.2844>
- KANDEL, D. B., HU, M.-C., GRIESLER, P. and WALL, M. (2017). Increases from 2002 to 2015 in prescription opioid overdose deaths in combination with other substances. *Drug Alcohol Depend.* **178** 501–511.
- KUANG, D., BRANTINGHAM, P. J. and BERTOZZI, A. L. (2017). Crime topic modeling. *Crime Sci.* **6** 12.
- LEE, D. D. and SEUNG, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems* 556–562.
- MARS, S. G., BOURGOIS, P., KARANDINOS, G., MONTERO, F. and CICCARONE, D. (2014). “Every ‘never’ I ever said came true”: Transitions from opioid pills to heroin injecting. *Int. J. Drug Policy* **25** 257–266.
- MEGHANI, S. H., BYUN, E. and GALLAGHER, R. M. (2012). Time to take stock: A meta-analysis and systematic review of analgesic treatment disparities for pain in the United States. *Pain Med.* **13** 150–174.
- MOHLER, G. (2013). Modeling and estimation of multi-source clustering in crime and security data. *Ann. Appl. Stat.* **7** 1525–1539. [MR3127957 https://doi.org/10.1214/13-AOAS647](https://doi.org/10.1214/13-AOAS647)
- MOHLER, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast.* **30** 491–497.
- MOHLER, G. and BRANTINGHAM, P. J. (2018). Privacy preserving, crowd sourced crime Hawkes processes. In *2018 International Workshop on Social Sensing (SocialSens)* 14–19. IEEE.
- MOHLER, G., CARTER, J. and RAJE, R. (2018). Improving social harm indices with a modulated Hawkes process. *Int. J. Forecast.* **34** 431–439.
- MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011a). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. [MR2816705 https://doi.org/10.1198/jasa.2011.ap09546](https://doi.org/10.1198/jasa.2011.ap09546)
- MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011b). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. [MR2816705 https://doi.org/10.1198/jasa.2011.ap09546](https://doi.org/10.1198/jasa.2011.ap09546)
- MOHLER, G. O., SHORT, M. B., MALINOWSKI, S., JOHNSON, M., TITA, G. E., BERTOZZI, A. L. and BRANTINGHAM, P. J. (2015). Randomized controlled field trials of predictive policing. *J. Amer. Statist. Assoc.* **110** 1399–1411. [MR3449035 https://doi.org/10.1080/01621459.2015.1077710](https://doi.org/10.1080/01621459.2015.1077710)
- PANDEY, R. and MOHLER, G. O. (2018). Evaluation of crime topic models: Topic coherence vs spatial crime concentration. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)* 76–78. IEEE.
- RAY, B. R., LOWDER, E. M., KIVISTO, A. J., PHALEN, P. and GIL, H. (2018). EMS naloxone administration as non-fatal opioid overdose surveillance: 6-year outcomes in Marion County, Indiana. *Addiction* **113** 2271–2279. <https://doi.org/10.1111/add.14426>
- RUDD, R. A., PAULOZZI, L. J., BAUER, M. J., BURLESON, R. W., CARLSON, R. E., DAO, D., DAVIS, J. W., DUDEK, J., EICHLER, B. A. et al. (2014). Increases in heroin overdose deaths—28 states, 2010 to 2012. *Morb. Mort. Wkly. Rep.* **63** 849.
- SAMPSON, R. J. and GROVES, W. B. (1989). Community structure and crime: Testing social-disorganization theory. *Am. J. Sociol.* **94** 774–802.
- SETH, P., SCHOLL, L., RUDD, R. A. and BACON, S. (2018). Overdose deaths involving opioids, cocaine, and psychostimulants—United States, 2015–2016. *Morb. Mort. Wkly. Rep.* **67** 349–358. <https://doi.org/10.15585/mmwr.mm6712a1>
- STEVENS, K., KEGELMEYER, P., ANDRZEJEWSKI, D. and BUTTLER, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* 952–961. Association for Computational Linguistics.
- STRICKLER, G. K., ZHANG, K., HALPIN, J. M., BOHNERT, A. S., BALDWIN, G. and KREINER, P. W. (2019). Effects of mandatory prescription drug monitoring program (PDMP) use laws on prescriber registration and use and on risky prescribing. *Drug Alcohol Depend.*

- TOWNSLEY, M., HOMEL, R. and CHASELING, J. (2003). Infectious burglaries. A test of the near repeat hypothesis. *Br. J. Criminol.* **43** 615–633.
- TSUI, K.-L., WONG, S. Y., JIANG, W. and LIN, C.-J. (2011). Recent research and developments in temporal and spatiotemporal surveillance for public health. *IEEE Trans. Reliab.* **60** 49–58.
- 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>
- WEISBURD, D. (2015). The law of crime concentration and the criminology of place. *Criminology* **53** 133–157.
- YU, H.-L., KOLOVOS, A., CHRISTAKOS, G., CHEN, J.-C., WARMERDAM, S. and DEV, B. (2007). Interactive spatiotemporal modeling of health systems: The SEKS-GUI framework. *Stoch. Environ. Res. Risk Assess.* **21** 555–572. [MR2380675](#) <https://doi.org/10.1007/s00477-007-0135-0>
- ZEOLI, A. M., PIZARRO, J. M., GRADY, S. C. and MELDE, C. (2014). Homicide as infectious disease: Using public health methods to investigate the diffusion of homicide. *Justice Q.* **31** 609–632.
- 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>

AN APPROXIMATE BEST PREDICTION APPROACH TO SMALL AREA ESTIMATION FOR SHEET AND RILL EROSION UNDER INFORMATIVE SAMPLING

BY EMILY BERG^{1,*} AND JAE-KWANG KIM^{1,†}

¹Iowa State University, *emilyb@iastate.edu; †jkim@iastate.edu

The National Resources Inventory, a longitudinal survey of characteristics related to natural resources and agriculture on nonfederal U.S. land, has increasingly received requests for substate estimates in recent years. We consider estimation of erosion in subdomains of the Boone-Raccoon River Watershed. This region is of interest for its proximity to intensively cropped areas as well as important waterbodies. The NRI application requires a small area prediction approach that can handle nonlinear relationships and appropriately incorporate survey weights that may have nontrivial relationships to the response variable. Because of the informative design, the conditional distribution required to define a standard empirical Bayes predictor is unknown. We develop a prediction approach that utilizes the approximate distribution of survey weighted score equations arising from a specified two-level superpopulation model. We apply the method to construct estimates of mean erosion in small watersheds. We investigate the robustness of the procedure to an assumption of a constant dispersion parameter and validate the properties of the procedure through simulation.

REFERENCES

- ASPAROUHOV, T. (2006). General multi-level modeling with sampling weights. *Comm. Statist. Theory Methods* **35** 439–460. [MR2274063](#) <https://doi.org/10.1080/03610920500476598>
- BATTESE, G. E., HARTER, R. M. and FULLER, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.* **83** 28–36.
- BERG, E. and CHANDRA, H. (2014). Small area prediction for a unit-level lognormal model. *Comput. Statist. Data Anal.* **78** 159–175. [MR3212164](#) <https://doi.org/10.1016/j.csda.2014.03.007>
- BERG, E., CHANDRA, H. and CHAMBERS, R. (2016). Small area estimation for lognormal data. In *Analysis of Poverty Data by Small Area Estimation* (M. Pratesi, ed.) 279–298.
- BERG, E. and KIM, J.-K. (2021). Supplement to “An approximate best prediction approach to small area estimation for sheet and rill erosion under informative sampling.” <https://doi.org/10.1214/20-AOAS1388SUPP>
- BERG, E. and LEE, D. (2019). Small area prediction of quantiles for zero-inflated data and an informative sample design. *Stat. Theory Relat. Fields* **3** 114–128. [MR4028310](#) <https://doi.org/10.1080/24754269.2019.1666243>
- BREIDT, F. J. and FULLER, W. A. (1999). Design of supplemented panel surveys with application to the National Resources Inventory. *J. Agric. Biol. Environ. Stat.* **4** 391–403. [MR1816375](#) <https://doi.org/10.2307/1400497>
- DEVILLE, J. C., SÄRNDAL, C. E. and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *J. Amer. Statist. Assoc.* **88** 1013–1020.
- DREASSI, E., PETRUCCI, A. and ROCCO, E. (2014). Small area estimation for semicontinuous skewed spatial data: An application to the grape wine production in Tuscany. *Biom. J.* **56** 141–156. [MR3152707](#) <https://doi.org/10.1002/bimj.201200271>
- ELASHOFF, M. and RYAN, L. (2004). An EM algorithm for estimating equations. *J. Comput. Graph. Statist.* **13** 48–65. [MR2044870](#) <https://doi.org/10.1198/1061860043092>
- ERCIULESCU, A. L. and FULLER, W. A. (2019). Bootstrap prediction intervals for small area means from unit-level nonlinear models. *J. Surv. Statist. Methodol.* **7** 309–333.
- 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](#)
- FULLER, W. A. (2011). *Sampling Statistics* (Vol. 560). Wiley, New York.

- GHOSH, M. and MAITI, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika* **91** 95–112. MR2050462 <https://doi.org/10.1093/biomet/91.1.95>
- JIANG, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *J. Statist. Plann. Inference* **111** 117–127. MR1955876 [https://doi.org/10.1016/S0378-3758\(02\)00293-8](https://doi.org/10.1016/S0378-3758(02)00293-8)
- JIANG, J. and LAHIRI, P. (2006). Mixed model prediction and small area estimation. *TEST* **15** 1–96. MR2252522 <https://doi.org/10.1007/BF02595419>
- KIM, J. K., PARK, S. and LEE, Y. (2017). Statistical inference using generalized linear mixed models under informative cluster sampling. *Canad. J. Statist.* **45** 479–497. MR3729982 <https://doi.org/10.1002/cjs.11339>
- LOHR, S. L. (2009). *Sampling: Design and Analysis*. Nelson Education.
- NANDRAM, B. and SEDRANSK, J. (2002). Bayesian predictive inference for the proportion of eroded land in a small area in Iowa. *Environ. Ecol. Stat.* **9** 221–236. MR1917877 <https://doi.org/10.1023/A:1016236809152>
- NEUHAUS, J. M. and MCCULLOCH, C. E. (2011). Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. *Biometrika* **98** 147–162. MR2804216 <https://doi.org/10.1093/biomet/asq066>
- 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.
- U.S. DEPARTMENT OF AGRICULTURE (2018). *Summary Report: 2015 National Resources Inventory, Natural Resources Conservation Service, Washington, DC, and Center for Survey Statistics and Methodology*. Iowa State Univ., Ames, IA. Available at <http://www.nrcs.usda.gov/technical/nri/15summary>.
- PFEFFERMANN, D. (2013). New important developments in small area estimation. *Statist. Sci.* **28** 40–68. MR3075338 <https://doi.org/10.1214/12-STS395>
- PFEFFERMANN, D. and SVERCHKOV, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *J. Amer. Statist. Assoc.* **102** 1427–1439. MR2412558 <https://doi.org/10.1198/016214507000001094>
- PRASAD, N. G. N. and RAO, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *J. Amer. Statist. Assoc.* **85** 163–171. MR1137362
- RAO, J. N. K. (2015). Inferential issues in model-based small area estimation: Some new developments. *Statist. Transit. New Ser.* **16** 491–510.
- 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>
- SÄRNDAL, C. E., SWENSSON, B. and WRETMAN, J. (2003). *Model Assisted Survey Sampling*. Springer, Berlin.
- SLUD, E. V. and MAITI, T. (2006). Mean-squared error estimation in transformed Fay–Herriot models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 239–257. MR2188984 <https://doi.org/10.1111/j.1467-9868.2006.00542.x>
- TORABI, M. and RAO, J. N. K. (2010). Mean squared error estimators of small area means using survey weights. *Canad. J. Statist.* **38** 598–608. MR2753004 <https://doi.org/10.1002/cjs.10078>
- VERRET, F., RAO, J. N. K. and HIDIROGLOU, M. A. (2015). Model-based small area estimation under informative sampling. *Survey Methodology* **41** 333–347.
- WISCHMEIR, W. H. and SMITH, D. D. (1965). Predicting rainfall erosion losses from cropland east of the Rocky Mountains: Guide for selection for practices for soil and water conservation. U.S. Department of Agriculture. Agricultural Handbook NO. 282.
- WOLTER, K. M. (2007). *Introduction to Variance Estimation*, 2nd ed. *Statistics for Social and Behavioral Sciences*. Springer, New York. MR2288542
- YI, G. Y., RAO, J. N. K. and LI, H. (2016). A weighted composite likelihood approach for analysis of survey data under two-level models. *Statist. Sinica* **26** 569–587. MR3497760
- YOU, Y. and RAO, J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canad. J. Statist.* **30** 431–439. MR1944372 <https://doi.org/10.2307/3316146>
- ZIMMERMANN, T. and MUNNICH, R. T. (2018). Small area estimation with a log-normal mixed model under informative sampling. *J. Off. Stat.* **34** 523–542.

ROBUST INFERENCE WHEN COMBINING INVERSE-PROBABILITY WEIGHTING AND MULTIPLE IMPUTATION TO ADDRESS MISSING DATA WITH APPLICATION TO AN ELECTRONIC HEALTH RECORDS-BASED STUDY OF BARIATRIC SURGERY

BY TANAYOTT THAWEETHAI¹, DAVID E. ARTERBURN², KAREN J. COLEMAN³ AND SEBASTIEN HANEUSE⁴

¹*Biostatistics Center, Massachusetts General Hospital, thaweethai@mgh.harvard.edu*

²*Health Research Institute, Kaiser Permanente Washington, David.E.Arterburn@kp.org*

³*Department of Research & Evaluation, Kaiser Permanente Southern California, Karen.J.Coleman@kp.org*

⁴*Department of Biostatistics, Harvard T. H. Chan School of Public Health, shaneuse@hsph.harvard.edu*

While electronic health records present a rich and promising data source for observational research, they are highly susceptible to missing data. For settings like these, Seaman et al. (*Biometrics* **68** (2012) 129–137) proposed a strategy wherein one handles missingness in some variables using inverse-probability weighting and others using multiple imputation. Seaman et al. (*Biometrics* **68** (2012) 129–137) show that Rubin’s variance estimator for averaging results across datasets is asymptotically valid when the analysis and imputation models are correctly specified and the weights are either known or correctly specified. Modeled after the approach of Robins and Wang (*Biometrika* **87** (2000) 113–124), we propose a method for asymptotically valid inference that is robust to violation of these conditions. Following a simulation study in which we demonstrate that a proposed variance estimator can reduce bias due to model misspecification, we illustrate this approach in an electronic health records-based study investigating whether differences in long-term weight loss between bariatric surgery techniques are associated with chronic kidney disease at baseline. We observe that the weight loss advantage after five years of Roux-en-Y gastric bypass surgery, compared to vertical sleeve gastrectomy, is less pronounced among patients with chronic kidney disease at baseline compared to those without.

REFERENCES

- ABADIE, A., IMBENS, G. W. and ZHENG, F. (2014). Inference for misspecified models with fixed regressors. *J. Amer. Statist. Assoc.* **109** 1601–1614. [MR3293613 https://doi.org/10.1080/01621459.2014.928218](https://doi.org/10.1080/01621459.2014.928218)
- ARTERBURN, D., LIVINGSTON, E. H., OLSEN, M. K., SMITH, V. A., KAVEE, A. L., KAHWATI, L. C., HENDERSON, W. G. and MACIEJEWSKI, M. L. (2013). Predictors of initial weight loss after gastric bypass surgery in twelve veterans affairs medical centers. *Obes. Res. Clin. Pract.* **7** e367–e376.
- ARTERBURN, D., WELLMAN, R., EMILIANO, A., SMITH, S. R., ODEGAARD, A. O., MURALI, S., WILLIAMS, N., COLEMAN, K. J., COURCOULAS, A. et al. (2018). Comparative effectiveness and safety of bariatric procedures for weight loss: A PCORnet cohort study. *Ann. Intern. Med.* **169** 741–750. <https://doi.org/10.7326/M17-2786>
- BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. [MR2216189 https://doi.org/10.1111/j.1541-0420.2005.00377.x](https://doi.org/10.1111/j.1541-0420.2005.00377.x)
- BRUCE BAYLEY, K., BELNAP, T., SAVITZ, L., MASICA, A., SHAH, N. and FLEMING, N. S. (2013). Challenges in using electronic health record data for CER: Experience of 4 learning organizations and solutions applied. *Med. Care* **51** S80–S86.
- BERNAARDS, C. A., BELIN, T. R. and SCHAFER, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat. Med.* **26** 1368–1382. [MR2345726 https://doi.org/10.1002/sim.2619](https://doi.org/10.1002/sim.2619)

Key words and phrases. Missing data, multiple imputation, inverse-probability weighting, model misspecification, electronic health records, bariatric surgery, obesity, chronic kidney disease.

- COURCOULAS, A. P., CHRISTIAN, N. J., O'ROURKE, R. W., DAKIN, G., PATCHEN DELLINGER, E., FLUM, D. R., MELISSA KALARCHIAN, P. D., MITCHELL, J. E., PATTERSON, E. et al. (2015). Preoperative factors and 3-year weight change in the Longitudinal Assessment of Bariatric Surgery (LABS) consortium. *Surg. Obes. Relat. Dis.* **11** 1109–1118.
- GAGNE, J. J., GLYNN, R. J., AVORN, J., LEVIN, R. and SCHNEEWEISS, S. (2011). A combined comorbidity score predicted mortality in elderly patients better than existing scores. *J. Clin. Epidemiol.* **64** 749–759.
- HANEUSE, S. (2016). Distinguishing selection bias and confounding bias in comparative effectiveness research. *Med. Care* **54** e23–e29. <https://doi.org/10.1097/MLR.0000000000000011>
- HANEUSE, S. and DANIELS, M. (2016). A general framework for considering selection bias in EHR-based studies: What data are observed and why? *EGEMs* **4** 1–17.
- HANEUSE, S. J. A. and SHORTREED, S. M. (2017). On the use of electronic health records. In *Methods in Comparative Effectiveness Research* 469–502. CRC Press, Boca Raton.
- HATOUM, I. J. and KAPLAN, L. M. (2012). Advantages of percent weight loss as a method of reporting weight loss after roux-en-y gastric bypass. *Obesity (Silver Spring, Md.)* **21**.
- HORTON, N. J. and KLEINMAN, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Amer. Statist.* **61** 79–90. [MR2339152 https://doi.org/10.1198/000313007X172556](https://doi.org/10.1198/000313007X172556)
- HUGHES, R. A., STERNE, J. A. C. and TILLING, K. (2016). Comparison of imputation variance estimators. *Stat. Methods Med. Res.* **25** 2541–2557. [MR3572869 https://doi.org/10.1177/0962280214526216](https://doi.org/10.1177/0962280214526216)
- HUQUE, M. H., CARLIN, J. B., SIMPSON, J. A. and LEE, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Med. Res. Methodol.* **18** 168–16.
- IMBENS, G. W. and KOLESÁR, M. (2016). Robust standard errors in small samples: Some practical advice. *Rev. Econ. Stat.* **98** 701–712.
- KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. [MR2420458 https://doi.org/10.1214/07-STS227](https://doi.org/10.1214/07-STS227)
- KENWARD, M. G. and CARPENTER, J. (2007). Multiple imputation: Current perspectives. *Stat. Methods Med. Res.* **16** 199–218. [MR2371006 https://doi.org/10.1177/0962280206075304](https://doi.org/10.1177/0962280206075304)
- KIM, J. K., BRICK, J. M., FULLER, W. A. and KALTON, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 509–521. [MR2278338 https://doi.org/10.1111/j.1467-9868.2006.00546.x](https://doi.org/10.1111/j.1467-9868.2006.00546.x)
- KROPKO, J., GOODRICH, B., GELMAN, A. and HILL, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Polit. Anal.* **22** 497–519.
- LEE, K. J. and CARLIN, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *Am. J. Epidemiol.* **171** 624–632.
- LEE, K. J. and CARLIN, J. B. (2017). Multiple imputation in the presence of non-normal data. *Stat. Med.* **36** 606–617. [MR3594613 https://doi.org/10.1002/sim.7173](https://doi.org/10.1002/sim.7173)
- LI, R. A., LIU, L., ARTERBURN, D., COLEMAN, K. J., COURCOULAS, A. P., FISHER, D., HANEUSE, S., JOHNSON, E., THEIS, M. K. et al. (2019). Five-year longitudinal cohort study of reinterventions after sleeve gastrectomy and roux-en-y gastric bypass. *Ann. Surg.* <https://doi.org/10.1097/sla.0000000000003401>
- LIESKE, J. C., MEHTA, R. A., MILLINER, D. S., RULE, A. D., BERGSTALH, E. J. and SARR, M. G. (2015). Kidney stones are common after bariatric surgery. *Kidney Inter., Suppl.* **87** 839–845. <https://doi.org/10.1038/ki.2014.352>
- MACIEJEWSKI, M. L., ARTERBURN, D. E., SCOYOC, L. V., SMITH, V. A., YANCY, W. S., WEIDENBACHER, H. J., LIVINGSTON, E. H. and OLSEN, M. K. (2016). Bariatric surgery and long-term durability of weight loss. *JAMA Surg.* **151** 1046–1055. <https://doi.org/10.1001/jamasurg.2016.2317>
- MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* **9** 538–558. <https://doi.org/10.1214/ss/1177010269>
- MENG, X.-L. and ROMERO, M. (2003). Discussion: Efficiency and self-efficiency with multiple imputation inference. *Int. Stat. Rev.* **71** 607–618.
- NASR, S. H., D'AGATI, V. D., SAID, S. M., STOKES, M. B., LARGOZA, M. V., RADHAKRISHNAN, J. and MARKOWITZ, G. S. (2008). Oxalate nephropathy complicating roux-en-y gastric bypass: An underrecognized cause of irreversible renal failure. *Clin. J. Am. Soc. Nephrol.* **3** 1676–1683. <https://doi.org/10.2215/CJN.02940608>
- NEFF, K. J., OLBERS, T. and LE ROUX, C. W. (2013). Bariatric surgery: The challenges with candidate selection, individualizing treatment and clinical outcomes. *BMC Med.* **11** 8. <https://doi.org/10.1186/1741-7015-11-8>
- NGUYEN, N. T., MASOOMI, H., LAUGENOUR, K., SANAIHA, Y., REAVIS, K. M., MILLS, S. D. and STAMOS, M. J. (2011). Predictive factors of mortality in bariatric surgery: Data from the nationwide inpatient sample. *Surgery* **150** 347–351.

- NIELSEN, S. F. (2003). Proper and improper multiple imputation. *Int. Stat. Rev.* **71** 593–607.
- PIVOVAROV, R., ALBERS, D. J., SEPULVEDA, J. L. and ELHADAD, N. (2014). Identifying and mitigating biases in EHR laboratory tests. *J. Biomed. Inform.* **51** 24–34. <https://doi.org/10.1016/j.jbi.2014.03.016>
- QUARTAGNO, M., CARPENTER, J. R. and GOLDSTEIN, H. (2019). Multiple imputation with survey weights: A multilevel approach. *J. Surv. Stat. Methodology*. smz036. <https://doi.org/10.1093/jssam/smz036>
- ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. [MR1294730](https://doi.org/10.1093/biomet/87.1.113)
- ROBINS, J. M. and WANG, N. (2000). Inference for imputation estimators. *Biometrika* **87** 113–124. [MR1766832](https://doi.org/10.1093/biomet/87.1.113)
- RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. [MR0455196](https://doi.org/10.1093/biomet/63.3.581)
- RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. Wiley Classics Library. Wiley Interscience, Hoboken, NJ. [MR2117498](https://doi.org/10.1201/9781439821862)
- SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Monographs on Statistics and Applied Probability **72**. CRC Press, London. [MR1692799](https://doi.org/10.1201/9781439821862)
- SCHENKER, N. and WELSH, A. H. (1988). Asymptotic results for multiple imputation. *Ann. Statist.* **16** 1550–1566. [MR0964938](https://doi.org/10.1214/aos/1176351053)
- SEAMAN, S. R. and VANSTEELANDT, S. (2018). Introduction to double robust methods for incomplete data. *Statist. Sci.* **33** 184–197. [MR3797709](https://doi.org/10.1214/18-STS647)
- SEAMAN, S. R. and WHITE, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Stat. Methods Med. Res.* **22** 278–295. [MR3190658](https://doi.org/10.1177/0962280210395740)
- SEAMAN, S. R., WHITE, I. R. and LEACY, F. P. (2014). Comment on “Analysis of longitudinal trials with protocol deviations: A framework for relevant, accessible assumptions, and inference via multiple imputation,” by Carpenter, Roger, and Kenward [MR3196115]. *J. Biopharm. Statist.* **24** 1358–1362. [MR3273793](https://doi.org/10.1080/10543406.2014.928306)
- SEAMAN, S. R., WHITE, I. R., COPAS, A. J. and LI, L. (2012). Combining multiple imputation and inverse-probability weighting. *Biometrics* **68** 129–137. [MR2909861](https://doi.org/10.1111/j.1541-0420.2011.01666.x)
- STEFANSKI, L. A. and BOOS, D. D. (2002). The calculus of M -estimation. *Amer. Statist.* **56** 29–38. [MR1939394](https://doi.org/10.1198/000313002753631330)
- TANG, Y. (2017). On the multiple imputation variance estimator for control-based and delta-adjusted pattern mixture models. *Biometrics* **73** 1379–1387. [MR3744550](https://doi.org/10.1111/biom.12702)
- THAWEETHAI, T., ARTERBURN, D. E., COLEMAN, K. J. and HANEUSE, S. (2021). Supplement to “Robust inference when combining inverse-probability weighting and multiple imputation to address missing data with application to an electronic health records-based study of bariatric surgery.” <https://doi.org/10.1214/20-AOAS1386SUPPA>, <https://doi.org/10.1214/20-AOAS1386SUPPB>
- TURGEON, N. A., PEREZ, S., MONDESTIN, M., DAVIS, S. S., LIN, E., TATA, S., KIRK, A. D., LARSEN, C. P., PEARSON, T. C. et al. (2012). The impact of renal function on outcomes of bariatric surgery. *J. Am. Soc. Nephrol.* **23** 885–894.
- VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). Mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** 1–67.
- VON HIPPEL, P. T. (2013). Should a normal imputation model be modified to impute skewed variables? *Sociol. Methods Res.* **42** 105–138. [MR3190726](https://doi.org/10.1177/0049124112464866)
- WANG, N. and ROBINS, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85** 935–948. [MR1666715](https://doi.org/10.1093/biomet/85.4.935)
- WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. [MR0640163](https://doi.org/10.2307/1912526)
- XIA, Y. and YANG, Y. (2016). Bias introduced by rounding in multiple imputation for ordered categorical variables. *Amer. Statist.* **70** 358–364. [MR3574788](https://doi.org/10.1080/00031305.2016.1200486)
- XU, D., DANIELS, M. J. and WINTERSTEIN, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics* **17** 589–602. [MR3603956](https://doi.org/10.1093/biostatistics/kxw009)
- YUCEL, R. M., HE, Y. and ZASLAVSKY, A. M. (2011). Gaussian-based routines to impute categorical variables in health surveys. *Stat. Med.* **30** 3447–3460. [MR2861625](https://doi.org/10.1002/sim.4355)
- ZELLMER, J. D., MATHIASON, M. A., KALLIES, K. J. and KOTHARI, S. N. (2014). Is laparoscopic sleeve gastrectomy a lower risk bariatric procedure compared with laparoscopic roux-en-y gastric bypass? A meta-analysis. *Am. J. Surg.* **208** 903–910.

SPIKE-AND-SLAB LASSO BICLUSTERING

BY GEMMA E. MORAN¹, VERONIKA ROČKOVÁ² AND EDWARD I. GEORGE³

¹Columbia Data Science Institute, Columbia University, gm2918@columbia.edu

²Booth School of Business, University of Chicago, Veronika.Rockova@ChicagoBooth.edu

³Department of Statistics, University of Pennsylvania, edgeorge@wharton.upenn.edu

Biclustering methods simultaneously group samples and their associated features. In this way, biclustering methods differ from traditional clustering methods, which utilize the entire set of features to distinguish groups of samples. Motivating applications for biclustering include genomics data, where the goal is to cluster patients or samples by their gene expression profiles; and recommender systems, which seek to group customers based on their product preferences. Biclusters of interest often manifest as rank-1 submatrices of the data matrix. This submatrix detection problem can be viewed as a factor analysis problem in which both the factors and loadings are sparse. In this paper, we propose a new biclustering method called Spike-and-Slab Lasso Biclustering (SSLB) which utilizes the Spike-and-Slab Lasso of Ročková and George (*J. Amer. Statist. Assoc.* **113** (2018) 431–444) to find such a sparse factorization of the data matrix. SSLB also incorporates an Indian Buffet Process prior to automatically choose the number of biclusters. Many biclustering methods make assumptions about the size of the latent biclusters; either assuming that the biclusters are all of the same size, or that the biclusters are very large or very small. In contrast, SSLB can adapt to find biclusters which have a continuum of sizes. SSLB is implemented via a fast EM algorithm with a variational step. In a variety of simulation settings, SSLB outperforms other biclustering methods. We apply SSLB to both a microarray dataset and a single-cell RNA-sequencing dataset and highlight that SSLB can recover biologically meaningful structures in the data. The SSLB software is available as an R/C++ package at <https://github.com/gemoran/SSLB>.

REFERENCES

- ARMAGAN, A., DUNSON, D. B. and CLYDE, M. (2011). Generalized beta mixtures of Gaussians. *Adv. Neural Inf. Process. Syst.* **24** 523–531.
- BAI, R., MORAN, G. E., ANTONELLI, J. L., CHEN, Y. and BOLAND, M. R. (2020). Spike-and-slab group Lasso for grouped regression and sparse generalized additive models. *J. Amer. Statist. Assoc.* 1–14. <https://doi.org/10.1080/01621459.2020.1765784>
- 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, Stat. Nonlinear Soft Matter Phys.* **67** 031902. <https://doi.org/10.1103/PhysRevE.67.031902>
- BHATTACHARYA, A. and CUI, Y. (2017). A GPU-accelerated algorithm for biclustering analysis and detection of condition-dependent coexpression network modules. *Sci. Rep.* **7** 4162. <https://doi.org/10.1038/s41598-017-04070-4>
- BOZDAĞ, D., KUMAR, A. S. and CATALYUREK, U. V. (2010). Comparative analysis of biclustering algorithms. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology* 265–274. ACM, New York.
- BOZDAĞ, D., PARVIN, J. D. and CATALYUREK, U. V. (2009). A biclustering method to discover co-regulated genes using diverse gene expression datasets. In *Bioinformatics and Computational Biology* 151–163. Springer, Berlin.

Key words and phrases. Bayes, biclustering, Spike-and-Slab Lasso, hierarchical modeling, variable selection, factor analysis.

- 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>
- CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874 <https://doi.org/10.1214/15-AOS1334>
- CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077 <https://doi.org/10.1214/12-AOS1029>
- CHENG, Y. and CHURCH, G. M. (2000). Biclustering of expression data. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* **8** 93–103.
- DENITTO, M., BICEGO, M., FARINELLI, A. and FIGUEIREDO, M. A. (2017). Spike and slab biclustering. *Pattern Recognit.* **72** 186–195.
- DESHPANDE, S. K., ROČKOVÁ, V. and GEORGE, E. I. (2019). Simultaneous variable and covariance selection with the multivariate spike-and-slab LASSO. *J. Comput. Graph. Statist.* **28** 921–931. MR4045858 <https://doi.org/10.1080/10618600.2019.1593179>
- DE CASTRO, P. A., DE FRANÇA, F. O., FERREIRA, H. M. and VON ZUBEN, F. J. (2007). Evaluating the performance of a biclustering algorithm applied to collaborative filtering—A comparative analysis. In *7th International Conference on Hybrid Intelligent Systems*, 2007. HIS 2007 65–70. IEEE, New York.
- DOSHI, F., MILLER, K., VAN GAEL, J. and TEH, Y. W. (2009). Variational inference for the Indian buffet process. In *Artificial Intelligence and Statistics* 137–144.
- EREN, K., DEVECİ, M., KÜÇÜKTUNÇ, O. and ÇATALYÜREK, Ü. V. (2012). A comparative analysis of biclustering algorithms for gene expression data. *Brief. Bioinform.* **14** 279–292.
- FAN, N., BOYKO, N. and PARDALOS, P. M. (2010). Recent advances of data biclustering with application in computational neuroscience. In *Computational Neuroscience. Springer Optim. Appl.* **38** 85–112. Springer, New York. MR2656809 https://doi.org/10.1007/978-0-387-88630-5_6
- FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications. Vol. II*, 2nd ed. Wiley, New York. MR0270403
- FRIEDMAN, J. H. and MEULMAN, J. J. (2004). Clustering objects on subsets of attributes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 815–849. MR2102467 <https://doi.org/10.1111/j.1467-9868.2004.02059.x>
- 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., MCDOWELL, I. C., ZHAO, S., BROWN, C. D. and ENGELHARDT, B. E. (2016). Context specific and differential gene co-expression networks via Bayesian biclustering. *PLoS Comput. Biol.* **12** e1004791.
- GHAHRAMANI, Z. and GRIFFITHS, T. L. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems* 475–482.
- GRIFFITHS, T. L. and GHAHRAMANI, Z. (2011). The Indian buffet process: An introduction and review. *J. Mach. Learn. Res.* **12** 1185–1224. MR2804598
- GU, J. and LIU, J. S. (2008). Bayesian biclustering of gene expression data. *BMC Genomics* **9** S4.
- HARTIGAN, J. A. (1972). Direct clustering of a data matrix. *J. Amer. Statist. Assoc.* **67** 123–129.
- HOCHREITER, S., BODENHOFER, U., HEUSEL, M., MAYR, A., MITTERECKER, A., KASIM, A., KHAMIAKOVA, T., VAN SANDEN, S., LIN, D. et al. (2010). FABIA: Factor analysis for bicluster acquisition. *Bioinformatics* **26** 1520–1527.
- HOWLADER, N., ALTEKRUSE, S. F., LI, C. I., CHEN, V. W., CLARKE, C. A., RIES, L. A. and CRONIN, K. A. (2014). US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status. *J. Natl. Cancer Inst.* **106** dju055.
- HUANG, M., WANG, J., TORRE, E., DUECK, H., SHAFFER, S., BONASIO, R., MURRAY, J. I., RAJ, A., LI, M. et al. (2018). SAVER: Gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15** 539.
- KAISER, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23** 187–200.
- KELLOKUMPU-LEHTINEN, P., TALPAZ, M., HARRIS, D., VAN, Q., KURZROCK, R. and ESTROV, Z. (1996). Leukemia-inhibitory factor stimulates breast, kidney and prostate cancer cell proliferation by paracrine and autocrine pathways. *Int. J. Cancer* **66** 515–519.
- KLUGER, Y., BASRI, R., CHANG, J. T. and GERSTEIN, M. (2003). Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Res.* **13** 703–716.
- LAZZERONI, L. and OWEN, A. (2002). Plaid models for gene expression data. *Statist. Sinica* **12** 61–86. MR1894189
- MADEIRA, S. C. and OLIVEIRA, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **1** 24–45.
- MAXWELL, J. C. (1860). V. Illustrations of the dynamical theory of gases.—Part I. On the motions and collisions of perfectly elastic spheres. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **19** 19–32.
- MORAN, G. E., ROČKOVÁ, V. and GEORGE, E. I. (2019). Variance prior forms for high-dimensional Bayesian variable selection. *Bayesian Anal.* **14** 1091–1119. MR4044847 <https://doi.org/10.1214/19-BA1149>

- MORAN, G. E., ROČKOVÁ, V. and GEORGE, E. I. (2021a). Supplement to “Spike-and-slab Lasso biclustering.” <https://doi.org/10.1214/20-AOAS1385SUPPA>
- MORAN, G. E., ROČKOVÁ, V. and GEORGE, E. I. (2021b). “Spike-and-slab Lasso biclustering” R package. <https://doi.org/10.1214/20-AOAS1385SUPPB>
- MORAN, G. E., ROČKOVÁ, V. and GEORGE, E. I. (2021c). Source code for “Spike-and-slab Lasso biclustering.” <https://doi.org/10.1214/20-AOAS1385SUPPC>
- MUCHERINO, A., PAPAJOJGI, P. J. and PARDALOS, P. M. (2009). *Data Mining in Agriculture. Springer Optimization and Its Applications* **34**. Springer, New York. MR2542502 <https://doi.org/10.1007/978-0-387-88615-2>
- ONITILLO, A. A., ENGEL, J. M., GREENLEE, R. T. and MUKESH, B. N. (2009). Breast cancer subtypes based on ER/PR and Her2 expression: Comparison of clinicopathologic features and survival. *Clin. Sedic. Res.* **7** 4–13.
- PADILHA, V. A. and CAMPELLO, R. J. G. B. (2017). A systematic comparative evaluation of biclustering techniques. *BMC Bioinform.* **18** 55. <https://doi.org/10.1186/s12859-017-1487-1>
- PEETERS, R. (2003). The maximum edge biclique problem is NP-complete. *Discrete Appl. Math.* **131** 651–654. MR2011374 [https://doi.org/10.1016/S0166-218X\(03\)00333-0](https://doi.org/10.1016/S0166-218X(03)00333-0)
- PRELIĆ, A., BLEULER, S., ZIMMERMANN, P., WILLE, A., BÜHLMANN, P., GRUISSEM, W., HENNIG, L., THIELE, L. and ZITZLER, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **22** 1122–1129.
- RANGAN, A. V., MCGROUTHER, C. C., KELSEY, J., SCHORK, N., STAHL, E., ZHU, Q., KRISHNAN, A., YAO, V., TROYANSKAYA, O. et al. (2018). A loop-counting method for covariate-corrected low-rank biclustering of gene-expression and genome-wide association study data. *PLoS Comput. Biol.* **14** e1006105.
- REYA, T., MORRISON, S. J., CLARKE, M. F. and WEISSMAN, I. L. (2001). Stem cells, cancer, and cancer stem cells. *Nature* **414** 105–111. <https://doi.org/10.1038/35102167>
- ROČKOVÁ, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Ann. Statist.* **46** 401–437. MR3766957 <https://doi.org/10.1214/17-AOS1554>
- 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>
- ROČKOVÁ, V. and GEORGE, E. I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.* **113** 431–444. MR3803476 <https://doi.org/10.1080/01621459.2016.1260469>
- ROHE, K. and ZENG, M. (2020). Vintage factor analysis with varimax performs statistical inference. Preprint. Available at [arXiv:2004.05387](https://arxiv.org/abs/2004.05387).
- SCHROEDER, M., HAIBE-KAINS, B., CULHANE, A., SOTIRIOU, C., BONTEMPI, G. and QUACKENBUSH, J. (2011). breastCancerNKI: Genexpression dataset published by van’t Veer et al. [2002] and van de Vijver et al. [2002] (NKI). R package version 1.12.0.
- SHABALIN, A. A., WEIGMAN, V. J., PEROU, C. M. and NOBEL, A. B. (2009). Finding large average submatrices in high dimensional data. *Ann. Appl. Stat.* **3** 985–1012. MR2750383 <https://doi.org/10.1214/09-AOAS239>
- TEH, Y. W., GRÜR, D. and GHAHRAMANI, Z. (2007). Stick-breaking construction for the Indian buffet process. In *Artificial Intelligence and Statistics* 556–563.
- VAN DE VIJVER, M. J., HE, Y. D., VAN’T VEER, L. J., DAI, H., HART, A. A., VOSKUIL, D. W., SCHREIBER, G. J., PETERSE, J. L., ROBERTS, C. et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347** 1999–2009.
- VAN’T VEER, L. J., DAI, H., VAN DE VIJVER, M. J., HE, Y. D., HART, A. A., MAO, M., PETERSE, H. L., VAN DER KOOP, K., MARTON, M. J. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415** 530.
- WHITFIELD, M. L., GEORGE, L. K., GRANT, G. D. and PEROU, C. M. (2006). Common markers of proliferation. *Nat. Rev. Cancer* **6** 99.
- YU, G., WANG, L.-G., HAN, Y. and HE, Q.-Y. (2012). clusterProfiler: An R package for comparing biological themes among gene clusters. *Omics. J. Integr. Biol.* **16** 284–287. <https://doi.org/10.1089/omi.2011.0118>
- ZEISEL, A., MUÑOZ-MANCHADO, A. B., CODELUPPI, S., LÖNNERBERG, P., LA MANNO, G., JURÉUS, A., MARQUES, S., MUNGUBA, H., HE, L. et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347** 1138–1142.
- ZHAN, T., RINDTORFF, N. and BOUTROS, M. (2017). Wnt signaling in cancer. *Oncogene* **36** 1461–1473. <https://doi.org/10.1038/onc.2016.304>
- ZHANG, M. H., MAN, H. T., ZHAO, X. D., DONG, N. and MA, S. L. (2014). Estrogen receptor-positive breast cancer molecular signatures and therapeutic potentials. *Biomed. Rep.* **2** 41–52.
- ZHU, Y., SHEN, X. and YE, C. (2016). Personalized prediction and sparsity pursuit in latent factor models. *J. Amer. Statist. Assoc.* **111** 241–252. MR3494656 <https://doi.org/10.1080/01621459.2014.999158>

RANDOM-EFFECTS META-ANALYSIS OF PHASE I DOSE-FINDING STUDIES USING STOCHASTIC PROCESS PRIORS

BY MORENO URSINO^{1,*}, CHRISTIAN RÖVER^{2,‡}, SARAH ZOHAR^{**1,†} AND
TIM FRIEDE^{**2,§}

¹Centre de Recherche des Cordeliers, INSERM, Sorbonne Université, USPC, Université de Paris, *moreno.ursino@inserm.fr;
[†]sarah.zohar@inserm.fr

²Department of Medical Statistics, University Medical Center Göttingen, ‡christian.roever@med.uni-goettingen.de;
§tim.friede@med.uni-goettingen.de

Phase I dose-finding studies aim at identifying the maximum tolerated dose (MTD). Often, several dose-finding studies are conducted with some variation in the administration mode or dose panel. For instance, sorafenib (BAY 43-900) was used as monotherapy in 36 phase I trials, according to a recent clinicaltrials.gov search. Since the toxicity may not be directly related to the specific indication, synthesizing the information from several studies might be worthwhile. However, this is rarely done in practice and only a fixed-effect meta-analysis framework was proposed to date. We developed a Bayesian random-effects meta-analysis methodology to pool several phase I trials and suggest the MTD. A curve free hierarchical model on the logistic scale with random effects, accounting for between-trial heterogeneity, is used to model the probability of toxicity across the investigated doses. An Ornstein–Uhlenbeck Gaussian process is adopted for the random effects structure. Prior distributions for the curve-free model are based on a latent Gamma process. An extensive simulation study showed good performance of the proposed method also under model deviations. Sharing information between phase I studies can improve the precision of MTD selection, at least when the number of trials is reasonably large.

REFERENCES

- ARROWSMITH, J. (2011). Trial watch: Phase III and submission failures: 2007–2010. *Nat. Rev. Drug. Discov.* **10** 87. <https://doi.org/10.1038/nrd3375>
- AWADA, A., HENDLISZ, A., GIL, T., BARTHOLOMEUS, S., MANO, M., DE VALERIOLA, D., STRUMBERG, D., BRENDL, E., HAASE, C. G. 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.
- BABB, J., ROGATKO, A. and ZACKS, S. (1998). Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Stat. Med.* **17** 1103–1120.
- BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions. The Theory and Application of Isotonic Regression*. Wiley, London-Sydney. MR0326887
- BORTHAKUR, G., KANTARJIAN, H., RAVANDI, F., ZHANG, W., KONOPLEVA, M., WRIGHT, J. J., FADERL, S., VERSTOVSEK, S., MATHEWS, S. et al. (2011). Phase I study of sorafenib in patients with refractory or relapsed acute leukemias. *Haematologica* **96** 62–68. <https://doi.org/10.3324/haematol.2010.030452>
- BRETZ, F., PINHEIRO, J. C. and BRANSON, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* **61** 738–748. MR2196162 <https://doi.org/10.1111/j.1541-0420.2005.00344.x>
- CHEN, Y.-B., LI, S., LANE, A. A., CONNOLLY, C., DEL RIO, C., VALLES, B., CURTIS, M., BALLEEN, K., CUTLER, C., DEY, B. R. et al. (2014). Phase I trial of maintenance sorafenib after allogeneic hematopoietic stem cell transplantation for fms-like tyrosine kinase 3 internal tandem duplication acute myeloid leukemia. *Biology of Blood and Marrow Transplantation* **20** 2042–2048.
- CHEUNG, Y. K. (2011). *Dose Finding by the Continual Reassessment Method*. Chapman and Hall/CRC, Boca Raton.

- CHEUNG, K. (2013). dform: Dose-finding by the continual reassessment method. R package version 0.2-2.
- CHEVRET, S., ed. (2006). *Statistical Methods for Dose-Finding Experiments. Statistics in Practice*. Wiley, Chichester. [MR2274231 https://doi.org/10.1002/0470861258](https://doi.org/10.1002/0470861258)
- CLARK, J. W., EDER, J. P., RYAN, D., LATHIA, C. and LENZ, H.-J. (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.
- COMETS, E. and ZOHAR, S. (2009). A survey of the way pharmacokinetics are reported in published phase I clinical trials, with an emphasis on oncology. *Clinical Pharmacokinetics* **48** 387–395.
- CRUMP, M., HEDLEY, D., KAMEL-REID, S., LEBER, B., WELLS, R., BRANDWEIN, J., BUCKSTEIN, R., KASSIS, J., MINDEN, M. et al. (2010). A randomized phase I clinical and biologic study of two schedules of sorafenib in patients with myelodysplastic syndrome or acute myeloid leukemia: A NCIC (National Cancer Institute of Canada) Clinical Trials Group Study. *Leukemia & Lymphoma* **51** 252–260.
- DOOB, J. L. (1942). The Brownian movement and stochastic equations. *Ann. of Math.* (2) **43** 351–369. [MR0006634 https://doi.org/10.2307/1968873](https://doi.org/10.2307/1968873)
- FRIEDE, T., RÖVER, C., WANDEL, S. and NEUENSCHWANDER, B. (2017). Meta-analysis of few small studies in orphan diseases. *Res. Synth. Methods* **8** 79–91. <https://doi.org/10.1002/jrsm.1217>
- FRIEDE, T., POSCH, M., ZOHAR, S., ALBERTI, C., BENDA, N., COMETS, E., DAY, S., DMITRIENKO, A., GRAF, A. et al. (2018). Recent advances in methodology for clinical trials in small populations: The InSPiRe project. *Orphanet Journal of Rare Diseases* **13** 186.
- FURUSE, J., ISHII, H., NAKACHI, K., SUZUKI, E., SHIMIZU, S. and NAKAJIMA, K. (2008). Phase I study of sorafenib in Japanese patients with hepatocellular carcinoma. *Cancer Science* **99** 159–165.
- GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2** 1360–1383. [MR2655663 https://doi.org/10.1214/08-AOAS191](https://doi.org/10.1214/08-AOAS191)
- 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](https://doi.org/10.1002/9781118414113)
- GOYA, H., KURAISHI, H., KOYAMA, S., ICHIYAMA, T., YOSHIIKE, F., HIRAI, K., AGATSUMA, T., TATEISHI, K., KANDA, S. et al. (2012). Phase I/II study of S-1 combined with biweekly irinotecan chemotherapy in previously treated advanced non-small cell lung cancer. *Cancer Chemotherapy and Pharmacology* **70** 691–697.
- GÜNHAN, B. K., RÖVER, S. and FRIEDE, T. (2020). Meta-analysis of few studies involving rare events. *Res. Synth. Methods* **11** 74–90.
- GÜNHAN, B. K., WEBER, S. and FRIEDE, T. (2020). A Bayesian time-to-event pharmacokinetic model for phase I dose-escalation trials with multiple schedules. *Stat. Med.* **39** 3986–4000. <https://doi.org/10.1002/sim.8703>
- HARRISON, R. K. (2016). Phase II and phase III failures: 2013–2015. *Nat. Rev. Drug. Discov.* **15** 817–818. <https://doi.org/10.1038/nrd.2016.184>
- INOKUCHI, M., YAMASHITA, T., YAMADA, H., KOJIMA, K., ICHIKAWA, W., NIHEI, Z., KAWANO, T. and SUGIHARA, K. (2006). Phase I/II study of S-1 combined with irinotecan for metastatic advanced gastric cancer. *Br. J. Cancer* **94** 1130.
- ISHIMOTO, O., ISHIDA, T., HONDA, Y., MUNAKATA, M. and SUGAWARA, S. (2009). Phase I study of daily S-1 combined with weekly irinotecan in patients with advanced non-small cell lung cancer. *International Journal of Clinical Oncology* **14** 43–47.
- JACKSON, D., LAW, M., STIJNEN, T., VIECHTBAUER, W. and WHITE, I. R. (2018). A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Stat. Med.* **37** 1059–1085. [MR3777959 https://doi.org/10.1002/sim.7588](https://doi.org/10.1002/sim.7588)
- JIA, N., LIU, I., HALLDORSON, J., CARITHERS, R., PERKINS, J., REYES, J., YEH, M., STOHR, E., RAO, S. et al. (2013). Phase I adjuvant trial of sorafenib in patients with hepatocellular carcinoma after orthotopic liver transplantation. *Anticancer Research* **33** 2797–2800.
- KIM, M.-O., WANG, X., LIU, C., DORRIS, K., FOULADI, M. and SONG, S. (2017). Random-effects meta-analysis for systematic reviews of phase I clinical trials: Rare events and missing data. *Res. Synth. Methods* **8** 124–135.
- KOMATSU, Y., YUKI, S., FUSE, N., KATO, T., MIYAGISHIMA, T., KUDO, M., KUNIEDA, Y., TATEYAMA, M., WAKAHAMA, O. et al. (2010). Phase I/2 clinical study of irinotecan and oral S-1 (IRIS) in patients with advanced gastric cancer. *Advances in Therapy* **27** 483–492.
- KUSABA, H., ESAKI, T., FUTAMI, K., TANAKA, S., FUJISHIMA, H., MITSUGI, K., SAKAI, K., ARIYAMA, H., TANAKA, R. et al. (2010). Phase I/II study of a 3-week cycle of irinotecan and S-1 in patients with advanced colorectal cancer. *Cancer Science* **101** 2591–2595.

- LAWLESS, J. and CROWDER, M. (2004). Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Anal.* **10** 213–227. [MR2086957](#) <https://doi.org/10.1023/B:LIDA.0000036389.14073.dd>
- LE TOURNEAU, C., LEE, J. J. and SIU, L. L. (2009). Dose escalation methods in phase I cancer clinical trials. *J. Natl. Cancer Inst.* **101** 708–720.
- MILLER, A. A., MURRY, D. J., OWZAR, K., HOLLIS, D. R., KENNEDY, E. B., ABOU-ALFA, G., DESAI, A., HWANG, J., VILLALONA-CALERO, M. A. et al. (2009). Phase I and pharmacokinetic study of sorafenib in patients with hepatic or renal dysfunction: CALGB 60301. *J. Clin. Oncol.* **27** 1800–1805. [https://doi.org/10.1200/JCO.2008.20.0931](#)
- MINAMI, H., KAWADA, K., EBI, H., KITAGAWA, K., KIM, Y.-I., ARAKI, K., MUKAI, H., TAHARA, M., NAKAJIMA, H. et al. (2008). Phase I and pharmacokinetic study of sorafenib, an oral multikinase inhibitor, in Japanese patients with advanced refractory solid tumors. *Cancer Science* **99** 1492–1498.
- MOORE, M., HIRTE, H., SIU, L., OZA, A., HOTTE, S., PETRENCIUC, O., CIHON, F., LATHIA, C. and SCHWARTZ, B. (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., 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>
- NABORS, L. B., SUPKO, J. G., ROSENFELD, M., CHAMBERLAIN, M., PHUPHANICH, S., BATCHELOR, T., DESIDERI, S., YE, X., WRIGHT, J. et al. (2011). Phase I trial of sorafenib in patients with recurrent or progressive malignant glioma. *Neuro-Oncol.* **13** 1324–1330. [https://doi.org/10.1093/neuonc/nor145](#)
- NAKAFUSA, Y., TANAKA, M., OHTSUKA, T., MIYOSHI, A., KOHYA, N., KITAJIMA, Y., SATO, S., MOCHINAGA, S., DOHI, S. et al. (2008). Phase I/II study of combination therapy with S-1 and CPT-11 for metastatic colorectal cancer. *Molecular Medicine Reports* **1** 925–930.
- NEAL, R. M. (1999). Regression and classification using Gaussian process priors. In *Bayesian Statistics*, 6 (Alcoceber, 1998) (J. M. Bernardo et al., eds.) 475–501. Oxford Univ. Press, New York. [MR1723510](#)
- 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., MATANO, A., TANG, Z., ROYCHOUDHURY, S., WANDEL, S. and BAILEY, S. (2015). A Bayesian industry approach to phase I combination trials in oncology. In *Statistical Methods in Drug Combination Studies* (W. Zhao and H. Yang, eds.) **6**, 95–135. Chapman & Hall/CRC, Boca Raton.
- 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>
- O’QUIGLEY, J. and ZOHAR, S. (2010). Retrospective robustness of the continual reassessment method. *J. Biopharm. Statist.* **20** 1013–1025. [MR2756534](#) <https://doi.org/10.1080/10543400903315732>
- OGATA, Y., SASATOMI, T., AKAGI, Y., ISHIBASHI, N., MORI, S. and SHIROUZU, K. (2009). Dosage escalation study of S-1 and irinotecan in metronomic chemotherapy against advanced colorectal cancer. *Kurume Med. J.* **56** 1–7. [https://doi.org/10.2739/kurumemedj.56.1](#)
- PRETORIUS, A. G. P. S. (2016). Phase III trial failures: Costly, but preventable. *Applied Clinical Trials* **25**.
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR2514435](#)
- SCHWINGHAMMER, T. L. and KROBOTH, P. D. (1988). Basic concepts in pharmacodynamic modeling. *J. Clin. Pharmacol.* **28** 388–394.
- SHIOZAWA, M., SUGANO, N., TSUCHIDA, K., MORINAGA, S., AKAIKE, M. and SUGIMASA, Y. (2009). A phase I study of combination therapy with S-1 and irinotecan (CPT-11) in patients with advanced colorectal cancer. *Journal of Cancer Research and Clinical Oncology* **135** 365–370.
- STEUTEL, F. W. (1979). Infinite divisibility in theory and practice. *Scand. J. Stat.* **6** 57–64. [MR0538596](#)
- STRUMBERG, D., RICHLI, H., HILGER, R. A., SCHLEUCHER, N., KORFEE, S., TEWES, M., FAGHIH, M., BRENDL, E., VOLIOTIS, D. 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.
- TAKIUCHI, H., NARAHARA, H., TSUJINAKA, T., GOTOH, M., KAWABE, S.-I., KATSU, K.-I., IISHI, H., TATSUTA, M., FUJITANI, K. et al. (2005). Phase I study of S-1 combined with irinotecan (CPT-11) in patients with advanced gastric cancer (OGSG 0002). *Japanese Journal of Clinical Oncology* **35** 520–525.
- THOMAS, N., SWEENEY, K. and SOMAYAJI, V. (2014). Meta-analysis of clinical dose-response in a large drug development portfolio. *Stat. Biopharm. Res.* **6** 302–317.
- TURNER, R. M., DAVEY, J., CLARKE, M. J., THOMPSON, S. G. and HIGGINS, J. P. T. (2012). Predicting the extent of heterogeneity in meta-analysis, using empirical data from the cochrane database of systematic reviews. *Int. J. Epidemiol.* **41** 818–827.

- UHLÉNBECK, G. E. and ORNSTEIN, L. S. (1930). On the theory of Brownian motion. *Phys. Rev.* **36** 823–841.
- URSINO, M. (2020). MADF. Available at <https://github.com/moreno-ursino/MADF/>.
- URSINO, M., YUAN, Y., ALBERTI, C., COMETS, E., FAVRAIS, G., FRIEDE, T., LENTZ, F., STALLARD, N. and ZOHAR, S. (2019). A dose finding design for seizure reduction in neonates. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 427–444. [MR3903002 https://doi.org/10.1111/rssc.12289](https://doi.org/10.1111/rssc.12289)
- URSINO, M., RÖVER, C., ZOHAR, S. and FRIEDE, T. (2021a). Supplement to “Random-effects meta-analysis of Phase I dose-finding studies using stochastic process priors.” <https://doi.org/10.1214/20-AOAS1390SUPPA>.
- URSINO, M., RÖVER, C., ZOHAR, S. and FRIEDE, T. (2021b). Supplement to “Random-effects meta-analysis of Phase I dose-finding studies using stochastic process priors.” <https://doi.org/10.1214/20-AOAS1390SUPPB>.
- YODA, S., SOEJIMA, K., YASUDA, H., NAOKI, K., KAWADA, I., WATANABE, H., NAKACHI, I., SATOMI, R., NAKAYAMA, S. et al. (2011). A phase I study of S-1 and irinotecan combination therapy in previously treated advanced non-small cell lung cancer patients. *Cancer Chemotherapy and Pharmacology* **67** 717–722.
- 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](https://doi.org/10.1002/sim.4121)
- ZOHAR, S., LIAN, Q., LEVY, V., CHEUNG, K., IVANOVA, A. and CHEVRET, S. (2008). Quality assessment of phase I dose-finding cancer trials: Proposal of a checklist. *Clin. Trials* **5** 478–485.

MODEL FREE ESTIMATION OF GRAPHICAL MODEL USING GENE EXPRESSION DATA

BY JENNY YANG¹, YANG LIU², YUFENG LIU³ AND WEI SUN⁴

¹Department of Biostatistics, University of North Carolina at Chapel Hill, jhyang87@gmail.com

²Department of Mathematics and Statistics, Wright State University, yang.liu@wright.edu

³Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina, yfliu@email.unc.edu

⁴Public Health Science Division, Fred Hutchinson Cancer Research Center, wsun@fredhutch.org

Graphical model is a powerful and popular approach to study high-dimensional omic data, such as genome-wide gene expression data. Nonlinear relations between genes are widely documented. However, partly due to sparsity of data points in high-dimensional space (i.e., curse of dimensionality) and computational challenges, most available methods construct graphical models by testing linear relations. We propose to address this challenge by a two-step approach: first, use a model-free approach to prioritize the neighborhood of each gene; then, apply a nonparametric conditional independence testing method to refine such neighborhood estimation. Our method, named as “mofreds” (MOdel FRee Estimation of DAG Skeletons), seeks to estimate the skeleton of a directed acyclic graph (DAG) by this two-step approach. We studied the theoretical properties of mofreds and evaluated its performance in extensive simulation settings. We found mofreds has substantially better performance than the state-of-the-art method which is designed to detect linear relations of Gaussian graphical models. We applied mofreds to analyze gene expression data of breast cancer patients from The Cancer Genome Atlas (TCGA). We found that it discovers nonlinear relationships among gene pairs that are missed by the Gaussian graphical model methods.

REFERENCES

- CHICKERING, D. M. (2003). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3** 507–554. [MR1991085 https://doi.org/10.1162/153244303321897717](https://doi.org/10.1162/153244303321897717)
- COLOMBO, D. and MAATHUIS, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15** 3741–3782. [MR3291411](https://doi.org/10.1162/153244303321897717)
- COOK, R. D. (2000). Save: A method for dimension reduction and graphics in regression. *Comm. Statist. Theory Methods* **29** 2109–2121. <https://doi.org/10.1080/03610920008832598>
- DE CAMPOS, C. P. and JI, Q. (2011). Efficient structure learning of Bayesian networks using constraints. *J. Mach. Learn. Res.* **12** 663–689. [MR2786909](https://doi.org/10.1162/153244303321897717)
- ERDŐS, P. and RÉNYI, A. (1959). On random graphs. I. *Publ. Math. Debrecen* **6** 290–297. [MR0120167](https://doi.org/10.1162/153244303321897717)
- FOYGEL, R. and DRTON, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems* 604–612.
- HA, M. J., SUN, W. and XIE, J. (2016). PenPC: A two-step approach to estimate the skeletons of high-dimensional directed acyclic graphs. *Biometrics* **72** 146–155. [MR3500583 https://doi.org/10.1111/biom.12415](https://doi.org/10.1111/biom.12415)
- HARRIS, N. and DRTON, M. (2013). PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.* **14** 3365–3383. [MR3144465](https://doi.org/10.1162/153244303321897717)
- HOEFFDING, W. (1948). A non-parametric test of independence. *Ann. Math. Stat.* **19** 546–557. [MR0029139 https://doi.org/10.1214/aoms/1177730150](https://doi.org/10.1214/aoms/1177730150)
- HORVATH, S. and DONG, J. (2008). Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.* **4** e1000117, 27. [MR2443399 https://doi.org/10.1371/journal.pcbi.1000117](https://doi.org/10.1371/journal.pcbi.1000117)
- IMHOF, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48** 419–426. [MR0137199 https://doi.org/10.1093/biomet/48.3-4.419](https://doi.org/10.1093/biomet/48.3-4.419)

- KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.
- KOLACZYK, E. D. and CSÁRDI, G. (2014). *Statistical Analysis of Network Data with R. Use R!* Springer, New York. [MR3288852 https://doi.org/10.1007/978-1-4939-0983-4](https://doi.org/10.1007/978-1-4939-0983-4)
- LAI, Y., WU, B., CHEN, L. and ZHAO, H. (2004). A statistical method for identifying differential gene–gene co-expression patterns. *Bioinformatics* **20** 3146–3155.
- LEMEIRE, J., MEGANCK, S. and CARTELLA, F. (2010). Robust independence-based causal structure learning in absence of adjacency faithfulness. In *Probabilistic Graphical Models* 169.
- LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. [MR1137117](https://doi.org/10.1080/01621459.1991.10476817)
- LI, K.-C. (2002). Genome-wide coexpression dynamics: Theory and application. *Proc. Natl. Acad. Sci. USA* **99** 16875–16880.
- LIU, J., SUN, W. and LIU, Y. (2019). Joint skeleton estimation of multiple directed acyclic graphs for heterogeneous population. *Biometrics* **75** 36–47. [MR3953705 https://doi.org/10.1111/biom.12941](https://doi.org/10.1111/biom.12941)
- MAATHUIS, M. H., KALISCH, M. and BÜHLMANN, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.* **37** 3133–3164. [MR2549555 https://doi.org/10.1214/09-AOS685](https://doi.org/10.1214/09-AOS685)
- PEARL, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. [MR2548166 https://doi.org/10.1017/CBO9780511803161](https://doi.org/10.1017/CBO9780511803161)
- PENG, J., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. [MR2541591 https://doi.org/10.1198/jasa.2009.0126](https://doi.org/10.1198/jasa.2009.0126)
- SONG, K. (2009). Testing conditional independence via Rosenblatt transforms. *Ann. Statist.* **37** 4011–4045. [MR2572451 https://doi.org/10.1214/09-AOS704](https://doi.org/10.1214/09-AOS704)
- SPIRITES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR1815675](https://doi.org/10.1017/CBO9780511518031)
- SUN, W. and LI, L. (2012). Multiple loci mapping via model-free variable selection. *Biometrics* **68** 12–22. [MR2909849 https://doi.org/10.1111/j.1541-0420.2011.01650.x](https://doi.org/10.1111/j.1541-0420.2011.01650.x)
- WILDING, G. E. and MUDHOLKAR, G. S. (2008). Empirical approximations for Hoeffding’s test of bivariate independence using two Weibull extensions. *Stat. Methodol.* **5** 160–170. [MR2424751 https://doi.org/10.1016/j.stamet.2007.07.002](https://doi.org/10.1016/j.stamet.2007.07.002)
- ZHANG, J. and SPIRITES, P. (2016). The three faces of faithfulness. *Synthese* **193** 1011–1027. [MR3489140 https://doi.org/10.1007/s11229-015-0673-9](https://doi.org/10.1007/s11229-015-0673-9)
- ZHANG, K., PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence* 804–813. AUAI Press.
- YANG, J., LIU, Y., LIU, Y. and SUN, W. (2021). Supplement to “Model free estimation of graphical model using gene expression data.” <https://doi.org/10.1214/20-AOAS1380SUPPA>, <https://doi.org/10.1214/20-AOAS1380SUPPB>

AN EVALUATION OF STATISTICAL METHODS FOR AGGREGATE PATTERNS OF REPLICATION FAILURE

BY JACOB M. SCHAUER¹, KAITLYN G. FITZGERALD^{2,*}, SARAH PEKO-SPICER^{2,†},
MENA C. R. WHALEN^{2,‡}, RITA ZEJNULLAHI^{2,§} AND LARRY V. HEDGES^{2,¶}

¹*Feinberg School of Medicine, Northwestern University, jms@northwestern.edu*

²*Department of Statistics, Northwestern University, *kgfitzgerald@u.northwestern.edu;*

†SarahPekoSpicer2020@u.northwestern.edu; ‡menawhalen2020@u.northwestern.edu;

§RitaZejnullahi2020@u.northwestern.edu; ¶l-hedges@northwestern.edu

Several programs of research have sought to assess the replicability of scientific findings in different fields, including economics and psychology. These programs attempt to replicate several findings and use the results to say something about large-scale patterns of replicability in a field. However, little work has been done to understand the analytic methods used to do this, including what they are assessing and what their statistical properties are. This article examines several methods that have been used to study patterns of replicability in the social sciences. We describe in concrete terms how each method operationalizes the idea of “replication” and examine various statistical properties, including bias, precision and statistical power. We find that some analytic methods rely on an operational definition of replication that can be misleading. Other methods involve more sound definitions of replication, but most of these have limitations, such as large bias and uncertainty or low power. The findings suggest that we should use caution interpreting the results of such analyses and that work on more accurate methods may be useful to future replication research efforts.

REFERENCES

- BAHADUR, R. R. (1960). Stochastic comparison of tests. *Ann. Math. Stat.* **31** 276–295. <https://doi.org/10.1214/aoms/1177705894>
- BISHARA, A. J. and HITTNER, J. B. (2015). Reducing bias and error in the correlation coefficient due to non-normality. *Educ. Psychol. Meas.* **75** 785–804. <https://doi.org/10.1177/0013164414557639>
- BOLLEN, K., CACIOPPO, J. T., KAPLAN, R. M., KROSNICK, J. A. and OLDS, J. L. (2015). Reproducibility, replicability, and generalization in the social, behavioral, and economic sciences. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. National Science Foundation, Arlington, VA.
- BORENSTEIN, M., HEDGES, L. V., HIGGINS, J. P. T. and ROTHSTEIN, H. R. (2009). *Introduction to Meta-Analysis*. Wiley-Blackwell, Oxford.
- BRANDT, M. J., IJZERMAN, H., DIJKSTERHUIS, A., FARACH, F. J., GELLER, J., GINER-SOROLLA, R., GRANGE, J. A., PERUGINI, M., SPIES, J. R. et al. (2014). The replication recipe: What makes for a convincing replication? *J. Exp. Soc. Psychol.* **50** 217–224.
- CAMERER, C. F., DREBER, A., FORSELL, E., HO, T.-H., HÜBER, J., JOHANNESSON, M., KIRCHLER, M., ALMENBERG, J., ALTMEJD, A. et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science* **351** 1433–1436.
- CAMERER, C. F., DREBER, A., HOLZMEISTER, F., HO, T.-H., HÜBER, J., JOHANNESSON, M., KIRCHLER, N. G., NOSEK, B. A. et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat. Hum. Behav.* **2** 637–644.
- COOPER, H. M. (2011). *Reporting Research in Psychology: How to Meet Journal Article Reporting Standards*. APA Books, Washington, DC.
- COOPER, H. M., HEDGES, L. V. and VALENTINE, J. (2009). *The Handbook of Research Synthesis and Meta-Analysis*, 2nd ed. The Russell Sage Foundation, New York.

- CUMMING, G., FIDLER, F., KALINOWSKI, P. and LAI, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Aust. J. Psychol.* **64** 138–146.
- DERSIMONIAN, R. and LAIRD, N. M. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* **7**.
- DIAZ-FRANCES, E. and RUBIO, F. J. (2013). On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables. *Statist. Papers* **54** 309–323. <https://doi.org/10.1007/s00362-012-0429-2>
- ETZ, A. and VANDEKERCKHOVE, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE* **11** e0149794. <https://doi.org/10.1371/journal.pone.0149794>
- FIELLER, E. C. (1932). The distribution of the index in a bivariate normal distribution. *Biometrika* **24** 3–4. <https://doi.org/10.1093/biomet/24.3-4.428>
- FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10** 507–521.
- FISHER, R. A. (1921). On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron* **1** 3–32.
- GARREN, S. T. (1998). Maximum likelihood estimation of the correlation coefficient in a bivariate normal model with missing data. *Statist. Probab. Lett.* **38** 281–288. [MR1629923 https://doi.org/10.1016/S0167-7152\(98\)00035-2](https://doi.org/10.1016/S0167-7152(98)00035-2)
- GEARY, R. C. (1930). The frequency distribution of the quotient of two normal variates. *J. R. Stat. Soc.* **93** 442–446. <https://doi.org/10.2307/2342070>
- GILBERT, D. T., KING, G., PETTIGREW, S. and WILSON, T. D. (2016). Comment on “Estimating the reproducibility of psychological science”. *Science* **351** 1037–1037.
- GLESER, L. J. and OLKIN, I. (1994). Stochastically dependent effect sizes. *The handbook of research synthesis*. H. Cooper & L. V. Hedges (Eds.). Russell Sage Foundation 339–355.
- HARTGERINK, C. H. J., WICHERTS, J. M. and VAN ASSEN, M. A. L. M. (2017). Too good to be false: Non-significant results revisited. *Collabra: Psychology* **3** 9.
- HAYYA, J., ARMSTRONG, D. and GRESSIS, N. (1975). A note on the ratio of two normally distributed variables. *Manage. Sci.* **21** 1338–1341. <https://doi.org/10.1287/mnsc.21.11.1338>
- HEDGES, L. V. and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, FL. [MR0798597](https://doi.org/10.1006/0195-0678(1985)90:1<1338::AID-0000000000000000>3.0.CO;2)
- HEDGES, L. V. and PIGOTT, T. D. (2001). The power of statistical tests in meta-analysis. *Psychol. Methods* **6** 203–217. <https://doi.org/10.1037/1082-989X.6.3.203>
- HEDGES, L. V. and SCHAUER, J. M. (2019a). More than one replication study is needed for unambiguous tests of replication. *J. Educ. Behav. Stat.* **44** 543–570.
- HEDGES, L. V. and SCHAUER, J. M. (2019b). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychol. Methods* **24** 557–570.
- HEDGES, L. V. and VEVEA, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychol. Methods* **3** 486–504. <https://doi.org/10.1037/1082-989X.3.4.486>
- HSUEH, H.-M., CHEN, J. J. and KODELL, R. L. (2003). Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. *J. Biopharm. Statist.* **13** 675–689.
- HUNG, H. M. J., O’NEILL, R. T., BAUER, P. and KÖHNE, K. (1997). The behavior of the *P*-value when the alternative hypothesis is true. *Biometrics* **53** 11–22. [MR1450180 https://doi.org/10.2307/2533093](https://doi.org/10.2307/2533093)
- KALAIAN, H. K. and RAUDENBUSH, S. W. (1986). A multivariate mixed linear model for meta-analysis. *Psychol. Methods* **1** 227–235. <https://doi.org/10.1037/1082-989X.1.3.227>
- KLEIN, R. A., RATLIFF, K. A., VIANELLO, M., ADAMS, R. B., BANÍK, Š., BERNSTEIN, M. J., BOCIAN, K., BRANDT, M. J., BROOKS, B. et al. (2014). Investigating variation in replicability: A “many labs” replication project. *Soc. Psychol.* **45** 142–152. <https://doi.org/10.1027/1864-9335/a000178>
- KLEIN, R. A., VIANELLO, M., HASSELMAN, F., ADAMS, B. G., ADAMS, R. B., ALPER, S. et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* **1** 443–490.
- KLEIN, R. A., COOK, C. L., EBERSOLE, C. R., VITIELLO, C. A., NOSEK, B. A., CHARTIER, C. R., CHRISTOPHERSON, C. D. et al. (2019). Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. Available at <https://psyarxiv.com/vef2c>.
- LAMBERT, D. and HALL, W. J. (1982). Asymptotic lognormality of *P*-values. *Ann. Statist.* **10** 44–64. [MR0642718](https://doi.org/10.1214/aos/1176342718)
- MARSAGLIA, G. (2006). Ratios of normal variables. *J. Stat. Softw.* **16** 1–10. <https://doi.org/10.18637/jss.v016.i04>
- MAXWELL, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychol. Methods* **9** 147–163.
- MUCHINSKY, P. M. (1996). The correction for attenuation. *Educ. Psychol. Meas.* **56** 63–75.

- OLKIN, I. and PRATT, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Ann. Math. Stat.* **29** 201–211. [MR0093854 https://doi.org/10.1214/aoms/1177706717](https://doi.org/10.1214/aoms/1177706717)
- OPEN SCIENCE COLLABORATION (2015). Estimating the reproducibility of psychological science. *Science* **349** aac4716.
- PATIL, P., PENG, R. D. and LEEK, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect. Psychol. Sci.* **11** 539–544. <https://doi.org/10.1177/17456916166646366>
- RILEY, R. D., HIGGINS, J. P. T. and DEEKS, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ* **342**. d549.
- SCHAUER, J. M. (2018). Statistical Methods for Assessing Replication: A Meta-Analytic Framework. Doctoral Dissertation, Northwestern Univ., Evanston, IL.
- SCHAUER, J. M., FITZGERALD, K. G., PEKO-SPICER, S., WHALEN, M. C. R., ZEJNULLAHI, R. and HEDGES, L. V. (2021). Supplement to “An evaluation of statistical methods for aggregate patterns of replication failure.” <https://doi.org/10.1214/20-AOAS1387SUPP>
- SCHWEINSBERG, M., MADAN, N., VIANELLO, M., SOMMER, S. A., JORDAN, J., TIERNEY, W., AWTREY, E., ZHU, L. L., DIERMEIER, D. et al. (2016). The pipeline project: Pre-publication independent replications of a single laboratory’s research pipeline. *J. Exp. Soc. Psychol.* **66** 55–67.
- SHAPIN, S. and SCHAFER, S. (1985). *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton Univ. Press, Princeton, NJ.
- SIMONSOHN, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychol. Sci.* **26** 559–569.
- SPEARMAN, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* **15** 72–101.
- SPEARMAN, C. (1910). Correlation calculated from faulty data. *Br. J. Psychol.* **3** 271–295.
- STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. [MR1924302 https://doi.org/10.1111/1467-9868.00346](https://doi.org/10.1111/1467-9868.00346)
- TAMHANE, A. C. and SHI, J. (2009). Parametric mixture models for estimating the proportion of true null hypotheses and adaptive control of FDR. In *Optimality. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **57** 304–325. IMS, Beachwood, OH. [MR2681678 https://doi.org/10.1214/09-LNMS5718](https://doi.org/10.1214/09-LNMS5718)
- VALENTINE, J. C., BIGLAN, A., BORUCH, R. F., CASTRO, F. G., COLLINS, L. M., FLAY, B. R., KELLAM, S., MOŚCICKI, E. K. and SCHINKE, S. P. (2011). Replication in prevention science. *Prev. Sci.* **12** 103–117.
- VAN AERT, R. C. M. and VAN ASSEN, M. A. L. M. (2017). Bayesian evaluation of effect size after replicating an original study. *PLoS ONE* **12** e0175302.
- VANKOV, I., BOWERS, J. and MUNAFÒ, M. R. (2014). On the persistence of low power in psychological science. *Q. J. Exp. Psychol.* **67** 1037–1040.
- VERONIKI, A. A., JACKSON, D., VIECHTBAUER, W., BENDER, R., BOWDEN, J., KNAPP, G., KUSS, O., HIGGINS, J. P. T., LANGAN, D. et al. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res. Synth. Methods* **7** 55–79.
- WASSERSTEIN, R. L. and LAZAR, N. A. (2016). The ASA’s statement on *p*-values: Context, process, and purpose [Editorial]. *Amer. Statist.* **70** 129–133. [MR3511040 https://doi.org/10.1080/00031305.2016.1154108](https://doi.org/10.1080/00031305.2016.1154108)
- WOOD, P. and RANDALL, D. (2018). How bad is the government’s science? *Wall St. J.* Available at <https://www.wsj.com/articles/how-bad-is-the-governments-science-1523915765>.
- YONG, E. (2016). The inevitable evolution of bad science. *The Atlantic*. Available at <https://www.theatlantic.com/science/archive/2016/09/the-inevitable-evolution-of-bad-science/500609/>.

LARGE-SCALE INFERENCE OF CORRELATION AMONG MIXED-TYPE BIOLOGICAL TRAITS WITH PHYLOGENETIC MULTIVARIATE PROBIT MODELS

BY ZHENYU ZHANG¹, AKIHIKO NISHIMURA², PAUL BASTIDE³, XIANG JI⁴,
REBECCA P. PAYNE⁵, PHILIP GOULDER⁶, PHILIPPE LEMEY⁷ AND MARC A. SUCHARD⁸

¹*Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, zyz606@ucla.edu*

²*Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, akihiko4@gmail.com*

³*IMAG, Université de Montpellier, CNRS, paul.bastide@umontpellier.fr*

⁴*Department of Mathematics, School of Science & Engineering, Tulane University, xji4@tulane.edu*

⁵*Translational and Clinical Research Institute, Newcastle University, rebecca.payne2@ncl.ac.uk*

⁶*Department of Paediatrics, University of Oxford, HIV Pathogenesis Programme, Doris Duke Medical Research Institute, University of KwaZulu-Natal, Ragon Institute of MGH, MIT and Harvard University, philip.goulder@paediatrics.ox.ac.uk*

⁷*Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, philippe.lemey@kuleuven.be*

⁸*Departments of Biomathematics, Biostatistics and Human Genetics, University of California, Los Angeles, msuchard@ucla.edu*

Inferring concerted changes among biological traits along an evolutionary history remains an important yet challenging problem. Besides adjusting for spurious correlation induced from the shared history, the task also requires sufficient flexibility and computational efficiency to incorporate multiple continuous and discrete traits as data size increases. To accomplish this, we jointly model mixed-type traits by assuming latent parameters for binary outcome dimensions at the tips of an unknown tree informed by molecular sequences. This gives rise to a phylogenetic multivariate probit model. With large sample sizes, posterior computation under this model is problematic, as it requires repeated sampling from a high-dimensional truncated normal distribution. Current best practices employ multiple-try rejection sampling that suffers from slow-mixing and a computational cost that scales quadratically in sample size. We develop a new inference approach that exploits: (1) the bouncy particle sampler (BPS) based on piecewise deterministic Markov processes to simultaneously sample all truncated normal dimensions, and (2) novel dynamic programming that reduces the cost of likelihood and gradient evaluations for BPS to linear in sample size. In an application with 535 HIV viruses and 24 traits that necessitates sampling from a 12,840-dimensional truncated normal, our method makes it possible to estimate the across-trait correlation and detect factors that affect the pathogen's capacity to cause disease. This inference framework is also applicable to a broader class of covariance structures beyond comparative biology.

REFERENCES

- ANDRIEU, C. and LIVINGSTONE, S. (2019). Peskun–Tierney ordering for Markov chain and process Monte Carlo: Beyond the reversible scenario. Preprint. Available at [arXiv:1906.06197](https://arxiv.org/abs/1906.06197).
- 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)
- BARBU, C. M., HONG, A., MANNE, J. M., SMALL, D. S., CALDERÓN, J. E. Q., SETHURAMAN, K., QUISPE-MACHACA, V., ANCCA-JUÁREZ, J., DEL CARPIO, J. G. C. et al. (2013). The effects of city streets on an urban disease vector. *PLoS Comput. Biol.* **9** e1002801.
- BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. [MR2806429](https://doi.org/10.1093/biomet/asr013)

Key words and phrases. Bayesian phylogenetics, probit models, bouncy particle sampler, dynamic programming, HIV evolution.

- BIERKENS, J. and DUNCAN, A. (2017). Limit theorems for the zig-zag process. *Adv. in Appl. Probab.* **49** 791–825. [MR3694318 https://doi.org/10.1017/apr.2017.22](https://doi.org/10.1017/apr.2017.22)
- BIERKENS, J., BOUCHARD-CÔTÉ, A., DOUCET, A., DUNCAN, A. B., FEARNEHEAD, P., LIENART, T., ROBERTS, G. and VOLLMER, S. J. (2018). Piecewise deterministic Markov processes for scalable Monte Carlo on restricted domains. *Statist. Probab. Lett.* **136** 148–154. [MR3806858 https://doi.org/10.1016/j.spl.2018.02.021](https://doi.org/10.1016/j.spl.2018.02.021)
- BOUCHARD-CÔTÉ, A., VOLLMER, S. J. and DOUCET, A. (2018). The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method. *J. Amer. Statist. Assoc.* **113** 855–867. [MR3832232 https://doi.org/10.1080/01621459.2017.1294075](https://doi.org/10.1080/01621459.2017.1294075)
- CHIB, S. and GREENBERG, E. (1998). Analysis of multivariate probit models. *Biometrika* **85** 347–361.
- CLARK, J. S., NEMERGUT, D., SEYEDNASROLLAH, B., TURNER, P. J. and ZHANG, S. (2017). Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecol. Monogr.* **87** 34–56.
- CRAWFORD, H., PRADO, J. G., LESLIE, A., HUÉ, S., HONEYBORNE, I., REDDY, S., VAN DER STOK, M., MNCUBE, Z., BRANDER, C. et al. (2007). Compensatory mutation partially restores fitness and delays reversion of escape mutation within the immunodominant HLA-B*5703-restricted Gag epitope in chronic human immunodeficiency virus type 1 infection. *J. Virol.* **81** 8346–51. <https://doi.org/10.1128/JVI.00465-07>
- CYBIS, G. B., SINSHEIMER, J. S., BEDFORD, T., MATHER, A. E., LEMEY, P. and SUCHARD, M. A. (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Ann. Appl. Stat.* **9** 969–991. [MR3371344 https://doi.org/10.1214/15-AOAS821](https://doi.org/10.1214/15-AOAS821)
- 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)
- DRAENERT, R., LE GALL, S., PFAFFEROTT, K. J., LESLIE, A. J., CHETTY, P., BRANDER, C., HOLMES, E. C., CHANG, S.-C., FEENEY, M. E. et al. (2004). Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. *J. Exp. Med.* **199** 905–915. <https://doi.org/10.1084/jem.20031982>
- DUNSON, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 355–366. [MR1749544 https://doi.org/10.1111/1467-9868.00236](https://doi.org/10.1111/1467-9868.00236)
- FEDOROV, V., WU, Y. and ZHANG, R. (2012). Optimal dose-finding designs with correlated continuous and discrete responses. *Stat. Med.* **31** 217–234. [MR2878850 https://doi.org/10.1002/sim.4388](https://doi.org/10.1002/sim.4388)
- FELSENSTEIN, J. (1985). Phylogenies and the comparative method. *Amer. Nat.* **125** 1–15.
- FELSENSTEIN, J. (2005). Using the quantitative genetic threshold model for inferences between and within species. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **360** 1427–1434.
- FELSENSTEIN, J. (2011). A comparative method for both discrete and continuous characters using the threshold model. *Amer. Nat.* **179** 145–156.
- GELMAN, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics* **48** 432–435. [MR2252307 https://doi.org/10.1198/004017005000000661](https://doi.org/10.1198/004017005000000661)
- GELMAN, A., RUBIN, D. B. et al. (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 https://doi.org/10.1002/9781118414113](https://doi.org/10.1002/9781118414113)
- GOULDER, P. J. and WALKER, B. D. (2012). HIV and HLA class I: An evolving relationship. *Immunity* **37** 426–440.
- GRAFEN, A. (1989). The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **326** 119–157.
- 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](https://doi.org/10.1007/s13253-018-00348-w)
- HEMELAAR, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends Mol. Med.* **18** 182–192. <https://doi.org/10.1016/j.molmed.2011.12.001>
- 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/10.1016/j.jmlr.2014.03.003](https://doi.org/10.1016/j.jmlr.2014.03.003)
- HUANG, A. and WAND, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Anal.* **8** 439–451. [MR3066948 https://doi.org/10.1214/13-BA815](https://doi.org/10.1214/13-BA815)
- HUANG, K.-H. G., GOEDHALS, D., CARLSON, J. M., BROCKMAN, M. A., MISHRA, S., BRUMME, Z. L., HICKLING, S., TANG, C. S., MIURA, T. et al. (2011). Progression to AIDS in South Africa is associated with both reverting and compensatory viral mutations. *PLoS ONE* **6** e19018.
- IRVINE, K. M., RODHOUSE, T. J. and KEREN, I. N. (2016). Extending ordinal regression with a latent zero-augmented beta distribution. *J. Agric. Biol. Environ. Stat.* **21** 619–640. [MR3576641 https://doi.org/10.1007/s13253-016-0265-2](https://doi.org/10.1007/s13253-016-0265-2)

- IVES, A. R. and GARLAND, T. (2010). Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.* **59** 9–26. <https://doi.org/10.1093/sysbio/syp074>
- 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](https://doi.org/10.1080/01621459.2015.1123632)
- KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* **13** 235–248. [MR0671034 https://doi.org/10.1016/0304-4149\(82\)90011-4](https://doi.org/10.1016/0304-4149(82)90011-4)
- LEWANDOWSKI, D., KUROWICKA, D. and JOE, H. (2009). Generating random correlation matrices based on vines and extended onion method. *J. Multivariate Anal.* **100** 1989–2001. [MR2543081 https://doi.org/10.1016/j.jmva.2009.04.008](https://doi.org/10.1016/j.jmva.2009.04.008)
- LEWIS, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50** 913–925.
- LIU, J. S., WONG, W. H. and KONG, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *J. Roy. Statist. Soc. Ser. B* **57** 157–169. [MR1325382](https://doi.org/10.2307/2346252)
- MARTINEZ-PICADO, J., PRADO, J. G., FRY, E. E., PFAFFEROTT, K., LESLIE, A., CHETTY, S., THOBAG-GALE, C., HONEYBORNE, I., CRAWFORD, H. et al. (2006). Fitness cost of escape mutations in p24 *gag* in association with control of human immunodeficiency virus type 1. *J. Virol.* **80** 3617–3623.
- 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.
- MRODE, R. A. (2014). *Linear Models for the Prediction of Animal Breeding Values*. Cabi.
- 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](https://doi.org/10.1080/01621459.2012.762328)
- 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/10.1201/9781439829639_11)
- NOMURA, S., HOSOYA, N., BRUMME, Z. L., BROCKMAN, M. A., KIKUCHI, T., KOGA, M., NAKAMURA, H., KOIBUCHI, T., FUJII, T. et al. (2013). Significant reductions in Gag-protease-mediated HIV-1 replication capacity during the course of the epidemic in Japan. *J. Virol.* **87** 1465–1476.
- PAGEL, M. (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond., B Biol. Sci.* **255** 37–45.
- PAYNE, R., MUENCHHOFF, M., MANN, J., ROBERTS, H. E., MATTHEWS, P., ADLAND, E., HEMPENSTALL, A., HUANG, K.-H., BROCKMAN, M. et al. (2014). Impact of HLA-driven HIV adaptation on virulence in populations of high HIV seroprevalence. *Proc. Natl. Acad. Sci. USA* **111** E5393–E5400.
- PETERS, E. A. J. F. and DE WIT, G. (2012). Rejection-free Monte Carlo sampling for general potentials. *Phys. Rev. E* **85** 026703.
- POURMOHAMAD, T. and LEE, H. K. H. (2016). Multivariate stochastic process models for correlated responses of mixed type. *Bayesian Anal.* **11** 797–820. [MR3498046 https://doi.org/10.1214/15-BA976](https://doi.org/10.1214/15-BA976)
- PRINCE, J. L., CLAIBORNE, D. T., CARLSON, J. M., SCHAEFER, M., YU, T., LAHKE, S., PRENTICE, H. A., YUE, L., VISHWANATHAN, S. A. et al. (2012). Role of transmitted *gag* CTL polymorphisms in defining replicative capacity and early HIV-1 pathogenesis. *PLoS Pathog.* **8** e1003041.
- PYBUS, O. G., SUCHARD, M. A., LEMEY, P., BERNARDIN, F. J., RAMBAUT, A., CRAWFORD, F. W., GRAY, R. R., ARINAMINPATHY, N., STRAMER, S. L. et al. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci. USA* **109** 15066–15071.
- 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](https://doi.org/10.1111/j.1467-9868.2008.00700.x)
- SCHLIEP, E. M. and HOETING, J. A. (2013). Multilevel latent Gaussian process model for mixed discrete and continuous multivariate response data. *J. Agric. Biol. Environ. Stat.* **18** 492–513. [MR3142597 https://doi.org/10.1007/s13253-013-0136-z](https://doi.org/10.1007/s13253-013-0136-z)
- SONG, H., PAVLICEK, J. W., CAI, F., BHATTACHARYA, T., LI, H., IYER, S. S., BAR, K. J., DECKER, J. M., GOONETILLEKE, N. et al. (2012). Impact of immune escape mutations on HIV-1 fitness in the context of the cognate transmitted/founder genome. *Retrovirology* **9** 89.
- STAN DEVELOPMENT TEAM (2018). Stan Modeling Language Users Guide and Reference Manual, Version 2.18.0.
- SUCHARD, M. A., WEISS, R. E. and SINSHEIMER, J. S. (2001). Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* **18** 1001–1013.
- SUCHARD, M. A., LEMEY, P., BAELE, G., AYRES, D. L., DRUMMOND, A. J. and RAMBAUT, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4** vey016. <https://doi.org/10.1093/ve/vey016>
- TOKUDA, T., GOODRICH, B., VAN MECHELEN, I., GELMAN, A. and TUEBLINCKX, F. (2011). Visualizing distributions of covariance matrices. Tech. Rep., 18–18, Columbia Univ., New York, USA.

- TROYER, R. M., MCNEVIN, J., LIU, Y., ZHANG, S. C., KRIZAN, R. W., ABRAHA, A., TEBIT, D. M., ZHAO, H., AVILA, S. et al. (2009). Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response. *PLoS Pathog.* **5** e1000365. <https://doi.org/10.1371/journal.ppat.1000365>
- TUNG HO, L. S. and ANÉ, C. (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst. Biol.* **63** 397–408.
- VITEZICA, Z. G., VARONA, L. and LEGARRA, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* **195** 1223–1230.
- WANG, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Anal.* **7** 867–886. **MR3000017** <https://doi.org/10.1214/12-BA729>
- WRIGHT, S. (1934). An analysis of variability in number of digits in an inbred strain of Guinea pigs. *Genetics* **19** 506.
- WRIGHT, J. K., BRUMME, Z. L., CARLSON, J. M., HECKERMAN, D., KADIE, C. M., BRUMME, C. J., WANG, B., LOSINA, E., MIURA, T. et al. (2010). Gag-protease-mediated replication capacity in HIV-1 subtype C chronic infection: Associations with HLA type and clinical parameters. *J. Virol.* **84** 10820–10831.
- WRIGHT, J. K., NAIDOO, V. L., BRUMME, Z. L., PRINCE, J. L., CLAIBORNE, D. T., GOULDER, P. J., BROCKMAN, M. A., HUNTER, E. and NDUNG’U, T. (2012). Impact of HLA-B* 81-associated mutations in HIV-1 *gag* on viral replication capacity. *J. Virol.* **86** 3193–3199.
- ZHANG, Z., NISHIMURA, A., BASTIDE, P., JI, X., PAYNE, R. P., GOULDER, P., LEMEY, P. and SUCHARD, M. A. (2021). Supplement to “Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models.” <https://doi.org/10.1214/20-AOAS1394SUPP>

MONITORING VACCINE SAFETY BY STUDYING TEMPORAL VARIATION OF ADVERSE EVENTS USING VACCINE ADVERSE EVENT REPORTING SYSTEM

BY JING HUANG^{1,*}, YI CAI², JINGCHENG DU^{3,¶}, RUOSHA LI⁴,
SUSAN S. ELLENBERG^{1,†}, SEAN HENNESSY^{1,‡}, CUI TAO^{3,||} AND YONG CHEN^{1,§}

¹Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, jing14@upenn.edu; [†]sellenbe@penndmedicine.upenn.edu; [‡]hennessy@penndmedicine.upenn.edu; [§]yichen123@upenn.edu

²AT&T Chief Data Office, AT&T Services, Inc., yc010@att.com

³School of Biomedical Informatics, University of Texas Health Science Center at Houston, [¶]jingcheng.Du@uth.tmc.edu; ^{||}cui.Tao@uth.tmc.edu

⁴Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston, Ruosha.Li@uth.tmc.edu

The Vaccine Adverse Event Reporting System (VAERS) plays a vital role in vaccine safety surveillance. One of the main missions of VAERS is to monitor increases in reporting rate of adverse events, as such signals can indicate safety issues caused by update of vaccines or change in vaccine practices. Existing methods can rarely be used to monitor the temporal variation of reporting adverse events. In this paper we propose a composite likelihood based variance component model to study the temporal variation of reporting adverse events using VAERS data. The proposed method is devised to identify safety signals by testing the heterogeneity of reporting rates of adverse events across years. The proposed method accounts for the well-known underreporting of adverse events and the zero-inflation observations in passive surveillance reporting systems. We applied the proposed method to VAERS reports of trivalent influenza virus vaccine and identified 14 adverse events with significantly heterogeneous reporting rates over years and two of them have increasing trend of reporting rates from 1990 to 2013. Our findings provide early warning signals that can be more rigorously investigated in future studies of the vaccine.

REFERENCES

- BATE, A., LINDQUIST, M., EDWARDS, I., OLSSON, S., ORRE, R., LANSNER, A. and DE FREITAS, R. M. (1998). A Bayesian neural network method for adverse drug reaction signal generation. *Eur. J. Clin. Pharmacol.* **54** 315–321.
- CHEN, R. T., HABER, P. and MULLEN, J. R. (1995). Surveillance of the safety of simultaneous administration of vaccines: The CDC experience. *Ann. N.Y. Acad. Sci.* **754** 309–320.
- CHEN, Y. and LIANG, K.-Y. (2010). On the asymptotic behaviour of the pseudolikelihood ratio test statistic with boundary problems. *Biometrika* **97** 603–620. [MR2672486 https://doi.org/10.1093/biomet/asq031](https://doi.org/10.1093/biomet/asq031)
- CHEN, Y., NING, Y., HONG, C., and WANG, S. (2014). Semiparametric tests for identifying differentially methylated loci with case-control designs using illumina arrays. *Genet. Epidemiol.* **38** 42–50. <https://doi.org/10.1002/gepi.21774>
- CHEN, Y., HONG, C., and RILEY, R. (2015). An alternative pseudolikelihood method for multivariate random-effects meta-analysis. *Stat. Med.* **34** 361–380. <https://doi.org/10.1002/sim.6350>
- CHEN, Y., NING, J., NING, Y., LIANG, K.-Y., and BANDEEN-ROCHE, K. (2017). On pseudolikelihood inference for semiparametric models with boundary problems. *Biometrika* **104** 165–179. <https://doi.org/10.1093/biomet/asw072>
- CHEN, R. T., RASTOGI, S. C., MULLEN, J. R., HAYES, S. W., COCHI, S. L., DONLON, J. A. and WASSILAK, S. G. (1994). The vaccine adverse event reporting system (VAERS). *Vaccine* **12** 542–550.
- CHEN, R. T., GLASSER, J. W., RHODES, P. H., DAVIS, R. L., BARLOW, W. E., THOMPSON, R. S., MULLOOLY, J. P., BLACK, S. B., SHINEFIELD, H. R. et al. (1997). Vaccine safety datalink project: A new tool for improving vaccine safety monitoring in the United States. *Pediatrics* **99** 765–773.

- CLAYTON, D. and KALDOR, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43** 671–681.
- COX, D. R. and REID, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91** 729–737. MR2090633 <https://doi.org/10.1093/biomet/91.3.729>
- DU, J., CAI, Y., CHEN, Y. and TAO, C. (2016). Trivalent influenza vaccine adverse symptoms analysis based on MedDRA terminology using VAERS data in 2011. *J. Biomed. Semant.* **7** 13.
- DUMOUCHEL, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Amer. Statist.* **53** 177–190.
- EBERTH, J. M., KLINE, K. N., MOSKOWITZ, D. A., MONTEALEGRE, J. R. and SCHEURER, M. E. (2014). The role of media and the Internet on vaccine adverse event reporting: A case study of human papillomavirus vaccination. *J. Adolesc. Health* **54** 289–295.
- ELLENBERG, S. S. and CHEN, R. T. (1997). The complicated task of monitoring vaccine safety. *Public Health Rep.* **112** 10.
- EVANS, S., WALLER, P. C. and DAVIS, S. (2001). Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol. Drug Saf.* **10** 483–486.
- CENTERS FOR DISEASE CONTROL AND PREVENTION (1991). Pertussis vaccination: Acellular pertussis vaccine for the fourth and fifth doses of the DTP series; update to supplementary ACIP statement. *Morb. Mort. Wkly. Rep.* **40** 881–882.
- CENTERS FOR DISEASE CONTROL AND PREVENTION (2016a). Seasonal influenza. Available at <http://www.cdc.gov/flu/about/qa/disease.htm>.
- CENTERS FOR DISEASE CONTROL AND PREVENTION (2016b). Seasonal influenza vaccine safety: A summary for clinicians. Available at https://www.cdc.gov/flu/professionals/vaccination/vaccine_safety.htm.
- CENTERS FOR DISEASE CONTROL AND PREVENTION (2017a). Estimated influenza illnesses, medical visits, hospitalizations and deaths averted by vaccination in the United States. Available at <https://www.cdc.gov/flu/about/disease/2015-16.htm>.
- CENTERS FOR DISEASE CONTROL AND PREVENTION (2017b). Vaccine effectiveness. Available at <https://www.cdc.gov/flu/about/qa/vaccineeffect.htm>.
- CENTERS FOR DISEASE CONTROL AND PREVENTION (2017c). Vaccine safety. Available at <https://www.cdc.gov/vaccinesafety/ensuringsafety/monitoring/vaers/index.html>.
- CENTERS FOR DISEASE CONTROL AND PREVENTION (2017d). Vaccine safety datalink (VSD). Available at <https://www.cdc.gov/vaccinesafety/ensuringsafety/monitoring/vsd/index.html>.
- GODAMBE, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Stat.* **31** 1208–1211. MR0123385 <https://doi.org/10.1214/aoms/1177705693>
- GOLDMAN, G. (2013). Comparison of VAERS fetal-loss reports during three consecutive influenza seasons: Was there a synergistic fetal toxicity associated with the two-vaccine 2009/2010 season? *Hum. Exp. Toxicol.* **32** 464–475.
- GREENWOOD, P. E. and NIKULIN, M. S. (1996). *A Guide to Chi-Squared Testing*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley, New York. MR1379800
- GROHSKOPF, L. A., SOKOLOW, L. Z., OLSEN, S. J., BRESEE, J. S., BRODER, K. R. and KARRON, R. A. (2015). Prevention and control of influenza with vaccines: Recommendations of the advisory committee on immunization practices, United States, 2015–16 influenza season. *Morb. Mort. Wkly. Rep.* **64** 818–825.
- HABER, P., DESTEFANO, F., ANGULO, F. J., ISKANDER, J., SHADOMY, S. V., WEINTRAUB, E. and CHEN, R. T. (2004). Guillain–Barré syndrome following influenza vaccination. *JAMA* **292** 2478–2481.
- HUANG, L., ZALKIKAR, J. and TIWARI, R. C. (2011). A likelihood ratio test based method for signal detection with application to FDA’s drug safety data. *J. Amer. Statist. Assoc.* **106** 1230–1241. MR2896832 <https://doi.org/10.1198/jasa.2011.ap10243>
- HUANG, L., ZHENG, D., ZALKIKAR, J. and TIWARI, R. (2017). Zero-inflated Poisson model based likelihood ratio test for drug safety signal detection. *Stat. Methods Med. Res.* **26** 471–488. MR3592736 <https://doi.org/10.1177/0962280214549590>
- HUANG, J., NING, Y., CAI, Y., LIANG, K.-Y. and CHEN, Y. (2020). Composite likelihood inference under boundary conditions. *Statist. Sinica* **30** 1005–1025.
- HUANG, J., CAI, Y., DU, J., LI, R., ELLENBERG, S. S., HENNESSY, S., TAO, C. and CHEN, Y. (2021). Supplement to “Monitoring vaccine safety by studying temporal variation of adverse events using vaccine adverse event reporting system.” <https://doi.org/10.1214/20-AOAS1393SUPP>
- IQBAL, S., LI, R., GARGIULLO, P. and VELLOZZI, C. (2015). Relationship between Guillain–Barré syndrome, influenza-related hospitalizations, and influenza vaccine coverage. *Vaccine* **33** 2045–2049.
- LASKY, T., TERRACCIANO, G. J., MAGDER, L., KOSKI, C. L., BALLESTEROS, M., NASH, D., CLARK, S., HABER, P., STOLLEY, P. D. et al. (1998). The Guillain–Barré syndrome and the 1992–1993 and 1993–1994 influenza vaccines. *N. Engl. J. Med.* **339** 1797–1802.

- LIEVANO, F., GALEA, S. A., THORNTON, M., WIEDMANN, R. T., MANOFF, S. B., TRAN, T. N., AMIN, M. A., SEMINACK, M. M., VAGIE, K. A. et al. (2012). Measles, mumps, and rubella virus vaccine (M-M-RTM II): A review of 32 years of clinical and postmarketing experience. *Vaccine* **30** 6918–6926.
- LINDSAY, B. G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes* (Ithaca, NY, 1987). *Contemp. Math.* **80** 221–239. Amer. Math. Soc., Providence, RI. [MR0999014 https://doi.org/10.1090/conm/080/999014](https://doi.org/10.1090/conm/080/999014)
- MCNEIL, M. M., GEE, J., WEINTRAUB, E. S., BELONGIA, E. A., LEE, G. M., GLANZ, J. M., NORDIN, J. D., KLEIN, N. P., BAXTER, R. et al. (2014). The vaccine safety datalink: Successes and challenges monitoring vaccine safety. *Vaccine* **32** 5390–5398.
- MEDDRA, M. (2017). Introductory Guide MedDRA Version 20.1. MedDRA Maintenance and Support Services Organization.
- PEARSON, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag.* **50** 157–175.
- ROTHMAN, K. J., LANES, S. and SACKS, S. T. (2004). The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiol. Drug Saf.* **13** 519–523.
- SHIMABUKURO, T. T., NGUYEN, M., MARTIN, D. and DEStEFANO, F. (2015). Safety monitoring in the vaccine adverse event reporting system (VAERS). *Vaccine* **33** 4398–4405.
- SINGLETON, J. A., LLOYD, J. C., MOOTREY, G. T., SALIVE, M. E., CHEN, R. T., ELLENBERG, S., RAS-TOGI, S., KRUEGER, C., BRAUN, M. et al. (1999). An overview of the vaccine adverse event reporting system (VAERS) as a surveillance system. *Vaccine* **17** 2908–2917.
- THOMPSON, K. M., GELLIN, B. G., HINMAN, A. R. and ORENSTEIN, W. A. (2018). The National Vaccine Advisory Committee at 30: Impact and opportunity. *Vaccine* **36** 1330–1344. <https://doi.org/10.1016/j.vaccine.2018.01.068>
- VARIN, C., REID, N. and FIRTH, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42. [MR2796852 https://doi.org/10.1007/s11464-011-9285-2](https://doi.org/10.1007/s11464-011-9285-2)
- VELLOZZI, C., BURWEN, D. R., DOBARDZIC, A., BALL, R., WALTON, K. and HABER, P. (2009). Safety of trivalent inactivated influenza vaccines in adults: Background for pandemic influenza vaccine safety monitoring. *Vaccine* **27** 2114–2120.
- WEBER, J. C. P. (1984). Epidemiology of adverse reactions to nonsteroidal anti-inflammatory drugs. In *Advances in Inflammation Research* 1–7.
- ZHAO, Y., YI, M. and TIWARI, R. C. (2018). Extended likelihood ratio test-based methods for signal detection in a drug class with application to FDA’s adverse event reporting system database. *Stat. Methods Med. Res.* **27** 876–890. [MR3767629 https://doi.org/10.1177/0962280216646678](https://doi.org/10.1177/0962280216646678)
- ZHOU, W., POOL, V., ISKANDER, J. K., ENGLISH-BULLARD, R., BALL, R., WISE, R. P., HABER, P., PLESS, R. P., MOOTREY, G. et al. (2003). Surveillance for safety after immunization: Vaccine adverse event reporting system (VAERS)—United States, 1991–2001. *MMWR Surveill. Summ.* **52** 1–24.

THE ROLE OF SCALE IN THE ESTIMATION OF CELL-TYPE PROPORTIONS

BY GREGORY J. HUNT¹ AND JOHANN A. GAGNON-BARTSCH²

¹Department of Mathematics, William & Mary, ghunt@wm.edu

²Department of Statistics, University of Michigan, johanngb@umich.edu

Complex tissues are composed of a large number of different types of cells, each involved in a multitude of biological processes. Consequently, an important component to understanding such processes is understanding the cell-type composition of the tissues. Estimating cell-type composition using high-throughput gene expression data is known as cell-type deconvolution. In this paper we first summarize the extensive deconvolution literature by identifying a common regression-like approach to deconvolution. We call this approach the unified deconvolution-as-regression (UDAR) framework. While methods that fall under this framework all use a similar model, they fit using data on different scales. Two popular scales for gene expression data are logarithmic and linear. Unfortunately, each of these scales has problems in the UDAR framework. Using log-scale gene expressions proposes a biologically implausible model and using linear-scale gene expressions will lead to statistically inefficient estimators. To explore ways to address these issues, in this paper we consider how deconvolution methods may use an adjusted model that is a hybrid of the two scales. In analysis on simulations as well as a collection of eleven real benchmark datasets, we find a prototypical hybrid-scale adjustment to the UDAR framework improves statistical efficiency and robustness. More broadly, we believe these hybrid-scale modeling principles may be incorporated into many existing deconvolution methods.

REFERENCES

- ABBAS, A. R., WOLSLEGEL, K., SESHASAYEE, D., MODRUSAN, Z. and CLARK, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE* **4**. <https://doi.org/10.1371/journal.pone.0006098>
- ALTBOUM, Z., STEUERMAN, Y., DAVID, E., BARNETT-ITZHAKI, Z., VALADARSKY, L., KEREN-SHAUL, H., MENINGHER, T., MENDELSON, E., MANDELBOIM, M. et al. (2014). Digital cell quantification identifies global immune cell dynamics during influenza infection. *Mol. Syst. Biol.* **10** 1–14. <https://doi.org/10.1002/msb.134947>
- AYANA, R., SINGH, S. and PATI, S. (2018). Deconvolution of human brain cell type transcriptomes unraveled microglia-specific potential biomarkers. *Front. Neurology* **9** 266. <https://doi.org/10.3389/fneur.2018.00266>
- BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation **3** 993–1022.
- BOUQUET, J., SOLOSKI, M. J., SWEL, A., CHEADLE, C., FEDERMAN, S., BILLAUD, J.-N. and REBMAN, A. W. (2016). Longitudinal transcriptome analysis reveals a sustained differential gene expression signature in patients treated for acute lyme disease **7** 1–11. <https://doi.org/10.1126/mBio.00100-16.Editor>
- CAPURRO, A., BODEA, L. G., SCHAEFER, P., LUTHI-CARTER, R. and PERREAU, V. M. (2015). Computational deconvolution of genome wide expression data from Parkinson’s and Huntington’s disease brain tissues using population-specific expression analysis. *Front. Neurosci.* **9** 1–12. <https://doi.org/10.3389/fnins.2015.00441>
- CONESA, A., MADRIGAL, P., TARAZONA, S., GOMEZ-CABRERO, D., CERVERA, A., MCPHERSON, A., SZCZEŚNIAK, M. W., GAFFNEY, D. J., ELO, L. L. et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17** 13. <https://doi.org/10.1186/s13059-016-0881-8>
- DU, R., CAREY, V. and WEISS, S. T. (2019). DeconvSeq: Deconvolution of cell mixture distribution in sequencing data. *Bioinformatics* **35** 5095–5102. <https://doi.org/10.1093/bioinformatics/btz444>
- EDGAR, R. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30** 207–210. <https://doi.org/10.1093/nar/30.1.207>

- FERNÁNDEZ, E., MAHMOUD, Y., VEIGAS, F., ROCHA, D., BALZARINI, M., LUJAN, H., RABINOVICH, G. and GIROTTI, M. R. (2019). MIXTURE: An improved algorithm for immune tumor microenvironment estimation based on gene expression data. *BioRxiv* 726562. <https://doi.org/10.1101/726562>
- FINOTELLO, F., MAYER, C., PLATTNER, C., LASCHNER, G., RIEDER, D., HACKL, H., KROGSDAM, A., LONCOVA, Z., POSCH, W. et al. (2019). Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Gen. Med.* **11** 34. <https://doi.org/10.1186/s13073-019-0638-6>
- GAUJOUX, R. (2013). An introduction to gene expression deconvolution and the CellMix package. 1–45.
- GEORGE, E. L. and PANOS, A. (2007). Does a high WBC count always signal infection? *Nursing* **37** 56hn15–56hn16. <https://doi.org/10.1097/01.NURSE.0000268785.73612.5c>
- GONG, T. and SZUSTAKOWSKI, J. D. (2013). DeconRNASeq: A statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-seq data. *Bioinformatics* **29** 1083–1085. <https://doi.org/10.1093/bioinformatics/btt090>
- GONG, T., HARTMANN, N., KOHANE, I. S., BRINKMANN, V., STAEDTLER, F., LETZKUS, M., BONGIOVANNI, S. and SZUSTAKOWSKI, J. D. (2011). Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS ONE* **6**. <https://doi.org/10.1371/journal.pone.0027156>
- HAGENAUER, M. H., LI, J. Z., WALSH, D. M., VAWTER, M. P., THOMPSON, R. C., TURNER, C. A., BUNNEY, W. E., MYERS, R. M., BARCHAS, J. D. et al. (2016). Inference of cell type composition from human brain transcriptomic datasets illuminates the effects of age, manner of death, dissection, and psychiatric diagnosis. *BioRxiv*.
- HARDIN, J. and WILSON, J. (2009). A note on oligonucleotide expression values not being normally distributed. *Biostatistics* **10** 446–450. <https://doi.org/10.1093/biostatistics/kxp003>
- HUNT, G. J. and GAGNON-BARTSCH, J. A. (2021). Supplement to “The Role of Scale in the Estimation of Cell-type Proportions.” <https://doi.org/10.1214/20-AOAS1395SUPPA>, <https://doi.org/10.1214/20-AOAS1395SUPPB>
- HUNT, G. J., FREYTAG, S., BAHLO, M. and GAGNON-BARTSCH, J. A. (2019). dtangle: Accurate and robust cell type deconvolution. *Bioinformatics* **35** 2093–2099. <https://doi.org/10.1093/bioinformatics/bty926>
- IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264. <https://doi.org/10.1093/biostatistics/4.2.249>
- SUN, X., SUN, S. and YANG, S. (2019). An efficient and flexible method for deconvoluting bulk RNA-seq data with single-cell RNA-seq data. *Cells* **8** 1161. <https://doi.org/10.3390/cells8101161>
- LIU, R., HOLIK, A. Z., SU, S., JANSZ, N., CHEN, K., LEONG, S., BLEWITT, M. E., SMYTH, G. K. and RITCHIE, M. E. (2015). Why weight ? Modelling sample and observational level variability improves power in RNA-seq analyses **43**. <https://doi.org/10.1093/nar/gkv412>
- LU, P., NAKORCHEVSKIY, A. and MARCOTTE, E. M. (2003). Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proc. Natl. Acad. Sci. USA* **100** 10370–10375. <https://doi.org/10.1073/pnas.1832361100>
- MOHAMMADI, S., ZUCKERMAN, N., GOLDSMITH, A. and GRAMA, A. (2015). A Critical Survey of Deconvolution Methods for Separating cell-types in Complex Tissues. arXiv 1–20.
- NEWMAN, A. M., LONG LIU, C., GREEN, M. R., GENTLES, A. J., FENG, W., XU, Y., HOANG, C. D., DIEHN, M. and ALIZADEH, A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12** 193–201. <https://doi.org/10.1016/j.molmed.2014.11.008>. Mitochondria
- PARSONS, J., MUNRO, S., PINE, P. S., MCDANIEL, J., MEHAFFEY, M. and SALIT, M. (2015). Using mixtures of biological samples as process controls for RNA-sequencing experiments. *BMC Genomics* 1–13. <https://doi.org/10.1186/s12864-015-1912-7>
- QIAO, W., QUON, G., CSASZAR, E., YU, M., MORRIS, Q. and ZANDSTRA, P. W. (2012). PERT: A method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput. Biol.* **8**. <https://doi.org/10.1371/journal.pcbi.1002838>
- RACLE, J., DE JONGE, K., BAUMGAERTNER, P., SPEISER, D. E. and GFELLER, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**. <https://doi.org/10.7554/eLife.26476>
- TU, Y., STOLOVITZKY, G. and KLEIN, U. (2002). Quantitative noise analysis for gene expression microarray experiments. *Proc. Natl. Acad. Sci. USA* **99** 14031–14036. <https://doi.org/10.1073/pnas.222164199>
- WANG, M., MASTER, S. R. and CHODOSH, L. A. (2006). Computational expression deconvolution in a complex mammalian organ. *BMC Bioinform.* **7** 328. <https://doi.org/10.1186/1471-2105-7-328>
- WANG, N., HOFFMAN, E. P., CHEN, L., CHEN, L., ZHANG, Z., LIU, C., YU, G., HERRINGTON, D. M., CLARKE, R. et al. (2016). Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci. Rep.* **6** 18909. <https://doi.org/10.1038/srep18909>

- WANG, Z., CAO, S., MORRIS, J. S., AHN, J., LIU, R., TYEKUCHEVA, S., GAO, F., LI, B., LU, W. et al. (2018). Transcriptome deconvolution of heterogeneous tumor samples with immune infiltration. *IScience* **9** 451–460. <https://doi.org/10.1016/j.isci.2018.10.028>
- WENG, L., DAI, H., ZHAN, Y., HE, Y., STEPANIANTS, S. B. and BASSETT, D. E. (2006). Rosetta error model for gene expression analysis. *Bioinformatics* **22** 1111–1121. <https://doi.org/10.1093/bioinformatics/btl045>
- WILSON, D. R., JIN, C., IBRAHIM, J. G. and SUN, W. (2020). ICeD-T provides accurate estimates of immune cell abundance in tumor samples by allowing for Aberrant gene expression patterns. *J. Amer. Statist. Assoc.* **115** 1055–1065. [MR4143449 https://doi.org/10.1080/01621459.2019.1654874](https://doi.org/10.1080/01621459.2019.1654874)
- ZWIENER, I., FRISCH, B. and BINDER, H. (2014). Transforming RNA-seq data to improve the performance of prognostic gene signatures. *PLoS ONE* **9** e85150. <https://doi.org/10.1371/journal.pone.0085150>

TWO-PHASE SAMPLING STRATEGIES FOR DESIGN-BASED MAPPING OF CONTINUOUS SPATIAL POPULATIONS IN ENVIRONMENTAL SURVEYS

BY LORENZO FATTORINI^{1,*}, SARA FRANCESCHI^{1,†}, MARZIA MARCHESELLI^{1,‡},
CATERINA PISANI^{1,§} AND LUCA PRATELLI²

¹Department of Economics and Statistics, University of Siena, *lorenzo.fattorini@unisi.it; †sara.franceschi@unisi.it;

‡marzia.marcheselli@unisi.it; §caterina.pisani@unisi.it

²Naval Academy, luca_pratelli@marina.difesa.it

The estimation of a surface throughout a continuum of points in a study area is addressed by means of two-phase sampling strategies. To this aim, a family of two-phase inverse distance weighting interpolators is introduced, and their design-based asymptotic properties are derived when the surface remains fixed and the number of sample points approaches infinity. In particular, conditions ensuring asymptotic unbiasedness and consistency are derived and are proven to hold for some of the most widely applied environmental sampling schemes. Furthermore, a computationally simple mean squared error estimator is proposed. Finally, a simulation study is performed to assess the theoretical results. The proposed strategy is adopted to provide the map of basal area in a forested region of Casentino Valley (Central Italy).

REFERENCES

- BAFFETTA, F., FATTORINI, L. and CORONA, P. (2011). Estimation of small woodlot and tree row attributes in large-scale forest inventories. *Environ. Ecol. Stat.* **18** 147–167. [MR2783687](#) <https://doi.org/10.1007/s10651-009-0125-0>
- BRUNO, F., COCCHI, D. and VAGHEGGINI, A. (2013). Finite population properties of individual predictors based on spatial patterns. *Environ. Ecol. Stat.* **20** 467–494. [MR3097174](#) <https://doi.org/10.1007/s10651-012-0229-9>
- CHIRICI, G., MCROBERTS, R. E., FATTORINI, L., MURA, M. and MARCHETTI, M. (2016). Comparing echo-based and canopy height model-based metrics for enhancing estimation of forest aboveground biomass in a model-assisted framework. *Remote Sens. Environ.* **174** 1–9. <https://doi.org/10.1016/j.rse.2015.11.010>
- CHIRICI, G., GIANNETTI, F., MCROBERTS, R. E., TRAVAGLINI, D., PECCHI, M., MASELLI, F., CHIESI, M. and CORONA, P. (2020). Wall-to-wall spatial prediction of growing stock volume based on Italian National Forest Inventory plots and remotely sensed data. *Int. J. Appl. Earth Obs. Geoinf.* **84** 101959. <https://doi.org/10.1016/j.jag.2019.101959>
- CORDY, C. B. (1993). An extension of the Horvitz–Thompson theorem to point sampling from a continuous universe. *Statist. Probab. Lett.* **18** 353–362. [MR1247446](#) [https://doi.org/10.1016/0167-7152\(93\)90028-H](https://doi.org/10.1016/0167-7152(93)90028-H)
- CORONA, P., FATTORINI, L., FRANCESCHI, S., SCRINZI, G. and TORRESAN, C. (2014). Estimation of standing wood volume in forest compartments by exploiting airborne laser scanning information: Model-based, design-based and hybrid perspectives. *Can. J. For. Res.* **44** 1303–1311. <https://doi.org/10.1139/cjfr-2014-0203>
- CORONA, P., FATTORINI, L., FRANCESCHI, S., MARCHESELLI, M., PISANI, C., CHIAVETTA, U. and PULETTI, N. (2019). Estimating tree diversity in forest ecosystems by two-phase inventories. *Environmetrics* **30** e2502, 13. [MR3919904](#) <https://doi.org/10.1002/env.2502>
- CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley, New York. [MR1239641](#) <https://doi.org/10.1002/9781119115151>
- DI BIASE, R. M., FATTORINI, L. and MARCHI, M. (2018). Statistical inferential techniques for approaching forest mapping. A review of methods. *Ann. Silv. Res.* **42** 46–58. <https://doi.org/10.12899/asr-1738>
- FAO (2011). Expert meeting on assessment of forest inventory approaches for REDD+. Meeting report. UN-REDD programme. Rome, 31 May and 1 June 2011. MRV Report 8. FAO, Rome.
- FATTORINI, L. (2015). Design-based methodological advances to support national forest inventories: A review of recent proposals. *IForest* **8** 6–11. <https://doi.org/10.3832/ifor1239-007>
- FATTORINI, L., FRANCESCHI, S. and PISANI, C. (2009). A two-phase sampling strategy for large-scale forest carbon budgets. *J. Statist. Plann. Inference* **139** 1045–1055. [MR2479848](#) <https://doi.org/10.1016/j.jspi.2008.06.014>

- FATTORINI, L., MARCHESELLI, M. and PISANI, C. (2006). A three-phase sampling strategy for large-scale multiresource forest inventories. *J. Agric. Biol. Environ. Stat.* **11** 296–316. <https://doi.org/10.1198/108571106X130548>
- FATTORINI, L., MARCHESELLI, M. and PRATELLI, L. (2018). Design-based maps for finite populations of spatial units. *J. Amer. Statist. Assoc.* **113** 686–697. [MR3832219 https://doi.org/10.1080/01621459.2016.1278174](https://doi.org/10.1080/01621459.2016.1278174)
- FATTORINI, L., MARCHESELLI, M., PISANI, C. and PRATELLI, L. (2017). Design-based asymptotics for two-phase sampling strategies in environmental surveys. *Biometrika* **104** 195–205. [MR3626472 https://doi.org/10.1093/biomet/asw062](https://doi.org/10.1093/biomet/asw062)
- FATTORINI, L., MARCHESELLI, M., PISANI, C. and PRATELLI, L. (2018). Design-based maps for continuous spatial populations. *Biometrika* **105** 419–429. [MR3804411 https://doi.org/10.1093/biomet/asy012](https://doi.org/10.1093/biomet/asy012)
- FATTORINI, L., DI BIASE, R. M., GIULIARELLI, D., MARCHESELLI, M., PISANI, C. and CORONA, P. (2019). Mapping the diversity of forest attributes: A design-based approach. *Can. J. For. Res.* **49** 190–197. <https://doi.org/10.1139/cjfr-2018-0204>
- FATTORINI, L., FRANCESCHI, S., MARCHESELLI, M., PISANI, C. and PRATELLI, L. (2021). Supplement to “Two-phase sampling strategies for design-based mapping of continuous spatial populations in environmental surveys.” <https://doi.org/10.1214/20-AOAS1392SUPP>
- GREGOIRE, T. G. (1998). Design-based and model-based inference in survey sampling: Appreciating the difference. *Can. J. For. Res.* **28** 1429–1447. <https://doi.org/10.1139/X98-166>
- GREGOIRE, T. G. and VALENTINE, H. T. (2008). *Sampling Strategies for Natural Resources and the Environment*. Chapman & Hall/CRC, Boca Raton, FL.
- ISAKI, C. T. and FULLER, W. A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.* <https://doi.org/10.2307/2287773> 89–96. [MR0648029 https://doi.org/10.2307/2287773](https://doi.org/10.2307/2287773)
- MANDALLAZ, D. (2008). *Sampling Techniques for Forest Inventories*. Chapman & Hall/CRC Applied Environmental Statistics. CRC Press, Boca Raton, FL. [MR2406002 MR2406002 https://doi.org/10.1201/9781584889779](https://doi.org/10.1201/9781584889779)
- OPSOMER, J. D., BREIDT, F. J., MOISEN, G. G. and KAUERMANN, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *J. Amer. Statist. Assoc.* **102** 400–409. [MR2370838 https://doi.org/10.1198/016214506000001491](https://doi.org/10.1198/016214506000001491)
- 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](https://doi.org/10.1007/978-1-4612-4378-6)
- TOBLER, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **46** 234–240. <https://doi.org/10.2307/143141>
- TOMPPA, L. M., GSCHWANTNER, R. E. and MCROBERTS, R. E. (2010). *National Forest Inventories: Pathways for Common Reporting*. Springer, Heidelberg. <https://doi.org/10.1007/978-90-481-3233-1>
- WALVOORT, D. J. J., BRUS, D. J. and DE GRUITER, J. J. (2010). An R package for spatial coverage sampling and random sampling from compact geographical strata by k -means. *Comput. Geosci.* **36** 1261–1267. <https://doi.org/10.1016/j.cageo.2010.04.005>

A REGRESSION DISCONTINUITY DESIGN FOR ORDINAL RUNNING VARIABLES: EVALUATING CENTRAL BANK PURCHASES OF CORPORATE BONDS

BY FAN LI¹, ANDREA MERCATANTI^{2,*}, TANELI MÄKINEN^{2,†} AND
ANDREA SILVESTRINI^{2,‡}

¹*Department of Statistical Science, Duke University, fl35@duke.edu*

²*DG for Economics, Statistics and Research, Bank of Italy* *andrea.mercatanti@bancaditalia.it;

†taneli.makinen@bancaditalia.it; ‡andrea.silvestrini@bancaditalia.it

Regression discontinuity (RD) is a widely used quasi-experimental design for causal inference. In the standard RD the assignment to treatment is determined by a continuous pretreatment variable (i.e., running variable) falling above or below a prefixed threshold. Recent applications increasingly feature ordered categorical or ordinal running variables which pose challenges to RD estimation due to the lack of a meaningful measure of distance. This paper proposes an RD approach for ordinal running variables under the local randomization framework. The proposal first estimates an ordered probit model for the ordinal running variable. The estimated probability of being assigned to treatment is then adopted as a latent continuous running variable and used to identify a covariate-balanced subsample around the threshold. Assuming local unconfoundedness of the treatment in the subsample, an estimate of the effect of the program is obtained by employing a weighted estimator of the average treatment effect. Two weighting estimators—overlap weights and ATT weights—as well as their augmented versions are considered. We apply the method to evaluate the causal effects of the corporate sector purchase programme (CSPP) of the European Central Bank which involves large-scale purchases of securities issued by corporations in the euro area. We find a statistically significant and negative effect of the CSPP on corporate bond spreads at issuance.

REFERENCES

- ABIDI, N. and FLORES, I. M. (2018). Who benefits from the corporate QE? A regression discontinuity design approach. ECB Working Paper Series No. 2145.
- AGRESTI, A. (2013). *Categorical Data Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. [MR3087436](#)
- ANG, J. S. and PATEL, K. A. (1975). Bond rating methods: Comparison and validation. *J. Finance* **30** 631–640.
- ANGRIST, J. D. and KRUEGER, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Q. J. Econ.* **106** 979–1014.
- ANGRIST, J. D. and LAVY, V. (1999). Using Maimonides’ rule to estimate the effect of class size on scholastic achievement. *Q. J. Econ.* **114** 533–575.
- ANGRIST, J. D. and ROKKANEN, M. (2012). Wanna get away? RD identification away from the cutoff. NBER Working Papers No. 18662, National Bureau of Economic Research.
- ANGRIST, J. D. and ROKKANEN, M. (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *J. Amer. Statist. Assoc.* **110** 1331–1344. [MR3449027](#) <https://doi.org/10.1080/01621459.2015.1012259>
- ARCE, O., GIMENO, R. and MAYORDOMO, S. (2017). Making room for the needy: The credit-reallocation effects of the ECB’s corporate QE. Documentos de Trabajo No. 1743, Banco de España.
- BECHT, M., POLO, A. and ROSSI, S. (2016). Does mandatory shareholder voting prevent bad acquisitions? *Rev. Financ. Stud.* **29** 3035–3067.

Key words and phrases. Asset purchase programs, augmented estimators, local unconfoundedness, M-estimation, ordered probit, regression discontinuity design, weighting.

- BERK, R. A. and DE LEEUW, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design. *J. Amer. Statist. Assoc.* **94** 1045–1052.
- BIAIS, B., DECLERCK, F., DOW, J., PORTES, R. and VON THADDEN, E.-L. (2006). European corporate bond markets: Transparency, liquidity, efficiency. CEPR Research Report, City of London.
- BLUME, M. E., LIM, F. and MACKINLAY, A. C. (1998). The declining credit quality of U.S. corporate debt: Myth or reality? *J. Finance* **53** 1389–1413.
- BRANSON, Z. and MEALLI, F. (2018). The local randomization framework for regression discontinuity designs: A review and some extensions. arXiv preprint [1810.02761](https://arxiv.org/abs/1810.02761).
- CALONICO, S., CATTANEO, M. D., FARRELL, M. H. and TITIUNIK, R. (2019). Regression discontinuity designs using covariates. *Rev. Econ. Stat.* **101** 442–451.
- CATTANEO, M. D., FRANDSEN, B. R. and TITIUNIK, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. senate. *J. Causal Inference* **3** 1–24.
- CATTANEO, M. D. and VAZQUEZ-BARE, G. (2016). The choice of neighborhood in regression discontinuity designs. *Observational Studies* **2** 134–146.
- CHAVA, S. and ROBERTS, M. R. (2008). How does financing impact investment? The role of debt covenants. *J. Finance* **63** 2085–2121.
- DICK-NIELSEN, J., FELDHÜTTER, P. and LANDO, D. (2012). Corporate bond liquidity before and after the onset of the subprime crisis. *J. Financ. Econ.* **103** 471–492.
- DONG, Y. (2015). Regression discontinuity applications with rounding errors in the running variable. *J. Appl. Econometrics* **30** 422–446. [MR3340946 https://doi.org/10.1002/jae.2369](https://doi.org/10.1002/jae.2369)
- GALEMA, R. and LUGO, S. (2017). When central banks buy corporate bonds: Target selection and impact of the European corporate sector purchase program. Discussion Paper Series No. 17-16, Utrecht Univ. School of Economics.
- GELMAN, A. and IMBENS, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *J. Bus. Econom. Statist.* **37** 447–456. [MR3968384 https://doi.org/10.1080/07350015.2017.1366909](https://doi.org/10.1080/07350015.2017.1366909)
- GENELETTI, S., O'KEEFE, A. G., SHARPLES, L. D., RICHARDSON, S. and BAIO, G. (2015). Bayesian regression discontinuity designs: Incorporating clinical knowledge in the causal analysis of primary care data. *Stat. Med.* **34** 2334–2352. [MR3358620 https://doi.org/10.1002/sim.6486](https://doi.org/10.1002/sim.6486)
- GREENSTONE, M. and GALLAGHER, J. (2008). Does hazardous waste matter? Evidence from the housing market and the superfund program. *Q. J. Econ.* **123** 951–1003.
- GROSSE-RUESCHKAMP, B., STEFFEN, S. and STREITZ, D. (2019). A capital structure channel of monetary policy. *J. Financ. Econ.* **133** 357–378.
- GÜNDÜZ, Y., OTTONELLO, G., PELIZZON, L., SCHNEIDER, M. and SUBRAHMANYAM, M. G. (2018). Lighting up the dark: A preliminary analysis of liquidity in the German corporate bond market. SAFE Working Paper No. 230, Leibniz Institute for Financial Research SAFE.
- HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331. [MR1612242 https://doi.org/10.2307/2998560](https://doi.org/10.2307/2998560)
- HAHN, J., TODD, P. and VAN DER KLAUW, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69** 201–209.
- HICKMAN, W. B. (1958). *Corporate Bond Quality and Investor Experience*. Princeton Univ. Press, Princeton, NJ.
- HIRANO, K. and IMBENS, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Serv. Outcomes Res. Methodol.* **2** 259–278.
- HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. [MR1995826 https://doi.org/10.1111/1468-0262.00442](https://doi.org/10.1111/1468-0262.00442)
- HJALMARSSON, R. (2009). Juvenile jails: A path to the straight and narrow or to hardened criminality? *J. Law Econ.* **52** 779–809.
- IMBENS, G. W. and LEMIEUX, T. (2008). Regression discontinuity designs: A guide to practice. *J. Econometrics* **142** 615–635. [MR2416821 https://doi.org/10.1016/j.jeconom.2007.05.001](https://doi.org/10.1016/j.jeconom.2007.05.001)
- IMBENS, G. and VAN DER KLAUW, W. (1995). Evaluating the cost of conscription in The Netherlands. *J. Bus. Econom. Statist.* **13** 207–215.
- IMBENS, G. W. and WAGER, S. (2019). Optimized regression discontinuity designs. *Rev. Econ. Stat.* **101** 264–278.
- KAO, C. and WU, C. (1990). Two-step estimation of linear models with ordinal unobserved variables: The case of corporate bonds. *J. Bus. Econom. Statist.* **8** 317–325.
- KAPLAN, R. S. and URWITZ, G. (1979). Statistical models of bond ratings: A methodological inquiry. *J. Bus.* **52** 231–261.
- KEYS, B. J., MUKHERJEE, T., SERU, A. and VIG, V. (2010). Did securitization lead to lax screening? Evidence from subprime loans. *Q. J. Econ.* **125** 307–362.

- KOLESÁR, M. and ROTHE, C. (2018). Inference in regression discontinuity designs with a discrete running variable. *Am. Econ. Rev.* **108** 2277–2304.
- KRAUS, A. and LITZENBERGER, R. H. (1973). A state-preference model of optimal financial leverage. *J. Finance* **28** 911–922.
- LEE, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *J. Econometrics* **142** 675–697. [MR2416824 https://doi.org/10.1016/j.jeconom.2007.05.004](https://doi.org/10.1016/j.jeconom.2007.05.004)
- LEE, D. S. and CARD, D. (2008). Regression discontinuity inference with specification error. *J. Econometrics* **142** 655–674. [MR2416823 https://doi.org/10.1016/j.jeconom.2007.05.003](https://doi.org/10.1016/j.jeconom.2007.05.003)
- LEE, D. S. and LEMIEUX, T. (2010). Regression discontinuity designs in economics. *J. Econ. Lit.* **48** 281–355.
- LI, F., MATTEI, A. and MEALLI, F. (2015). Evaluating the causal effect of university grants on student dropout: Evidence from a regression discontinuity design using principal stratification. *Ann. Appl. Stat.* **9** 1906–1931. [MR3456358 https://doi.org/10.1214/15-AOAS881](https://doi.org/10.1214/15-AOAS881)
- LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. [MR3803473 https://doi.org/10.1080/01621459.2016.1260466](https://doi.org/10.1080/01621459.2016.1260466)
- LI, F., ZASLAVSKY, A. M. and LANDRUM, M. B. (2013). Propensity score weighting with multilevel data. *Stat. Med.* **32** 3373–3387. [MR3074363 https://doi.org/10.1002/sim.5786](https://doi.org/10.1002/sim.5786)
- LI, F., MERCATANTI, A., MÄKINEN, T. and SILVESTRINI, A. (2021). Supplement to “A regression discontinuity design for ordinal running variables: Evaluating central bank purchases of corporate bonds.” <https://doi.org/10.1214/20-AOAS1396SUPP>
- LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **23** 2937–2960. <https://doi.org/10.1002/sim.1903>
- MATTEI, A. and MEALLI, F. (2016). Regression discontinuity designs as local randomized experiments. *Observational Studies* **2** 156–173.
- MEALLI, F. and RAMPICHINI, C. (2012). Evaluating the effects of university grants by using regression discontinuity designs. *J. Roy. Statist. Soc. Ser. A* **175** 775–798. [MR2948374 https://doi.org/10.1111/j.1467-985X.2011.01022.x](https://doi.org/10.1111/j.1467-985X.2011.01022.x)
- MERCATANTI, A. and LI, F. (2014). Do debit cards increase household spending? Evidence from a semi-parametric causal analysis of a survey. *Ann. Appl. Stat.* **8** 2485–2508. [MR3292506 https://doi.org/10.1214/14-AOAS784](https://doi.org/10.1214/14-AOAS784)
- MERTON, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *J. Finance* **29** 449–470.
- MIZEN, P. and TSOUKAS, S. (2012). Forecasting US bond default ratings allowing for previous and initial state dependence in an ordered probit model. *Int. J. Forecast.* **28** 273–287.
- MÄKINEN, T., LI, F., MERCATANTI, A. and SILVESTRINI, A. (2020). Effects of eligibility for central bank purchases on corporate bond spreads. Bank for International Settlements (BIS) Working Papers.
- PINCHES, G. E. and MINGO, K. A. (1973). A multivariate analysis of industrial bond ratings. *J. Finance* **28** 1–18.
- PINCHES, G. E. and MINGO, K. A. (1975). The role of subordination and industrial bond ratings. *J. Finance* **30** 201–206.
- POGUE, T. F. and SOLDOSKY, R. M. (1969). What’s in a bond rating. *J. Financ. Quant. Anal.* **4** 201–228.
- RAUH, J. D. (2006). Investment and financing constraints: Evidence from the funding of corporate pension plans. *J. Finance* **61** 33–71.
- 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. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.
- SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. With comments and a rejoinder by the authors. [MR1731478 https://doi.org/10.2307/2669923](https://doi.org/10.2307/2669923)
- SKOVRON, C. and TITIUNIK, R. (2015). A practical guide to regression discontinuity designs in political science. *Amer. J. Polit. Sci.* **2015** 1–36.
- STEFANSKI, L. A. and BOOS, D. D. (2002). The calculus of M -estimation. *Amer. Statist.* **56** 29–38. [MR1939394 https://doi.org/10.1198/000313002753631330](https://doi.org/10.1198/000313002753631330)
- THISTLETHWAITE, D. L. and CAMPBELL, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *J. Educ. Psychol.* **51** 309–317.
- VAN DER KLAUW, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *Internat. Econom. Rev.* **43** 1249–1287.
- 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)

- ZAGHINI, A. (2019). The CSPP at work: Yield heterogeneity and the portfolio rebalancing channel. *J. Corp. Finance* **56** 282–297.
- ZHOU, Y., MATSOUAKA, R. A. and THOMAS, L. (2020). Propensity score weighting under limited overlap and model misspecification. *Stat. Methods Med. Res.* **29** 3721–3756. MR4159223 <https://doi.org/10.1177/0962280220940334>

SPATIAL DISTRIBUTED LAG DATA FUSION FOR ESTIMATING AMBIENT AIR POLLUTION

BY JOSHUA L. WARREN¹, MARIE LYNN MIRANDA², JOSHUA L. TOOTOO^{3,*},
CLAIRE E. OSGOOD^{3,†} AND MICHELLE L. BELL⁴

¹Department of Biostatistics, Yale University, joshua.warren@yale.edu

²Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, mlm@nd.edu

³Children's Environmental Health Initiative, University of Notre Dame, *jtootoo@nd.edu; †cosgood@nd.edu

⁴School of Forestry and Environmental Studies, Department of Environmental Health Sciences, Yale University, michelle.bell@yale.edu

We introduce spatial (DLfuse) and spatiotemporal (DLfuseST) distributed lag data fusion methods for predicting point-level ambient air pollution concentrations, using, as input, gridded average pollution estimates from a deterministic numerical air quality model. The methods incorporate predictive information from grid cells surrounding the prediction location of interest and are shown to collapse to existing downscaling approaches when this information adds no benefit. The spatial lagged parameters are allowed to vary spatially/spatiotemporally to accommodate the setting where surrounding geographic information is useful in one area/time but not in another. We apply the new methods to predict ambient concentrations of eight-hour maximum ozone and 24-hour average PM_{2.5} at unobserved spatial locations and times, and compare the predictions with those from several state-of-the-art data fusion approaches. Results show that DLfuse and DLfuseST often provide improved model fit and predictive accuracy when the lagged information is shown to be beneficial. Code to apply the methods is available in the R package DLfuse.

REFERENCES

- BAEK, J., SÁNCHEZ, B. N., BERROCAL, V. J. and SANCHEZ-VAZNAUGH, E. V. (2016). Distributed lag models: Examining associations between the built environment and health. *Epidemiology* **27** 116.
- 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
- 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>
- BERMAN, J., JIN, L., BELL, M. and CURRIERO, F. C. (2019). Developing a geostatistical simulation method to inform the quantity and placement of new monitors for a follow-up air sampling campaign. *J. Expo. Sci. Environ. Epidemiol.* **29** 248.
- BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2010a). 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>
- BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2010b). A bivariate space-time downscaler under space and time misalignment. *Ann. Appl. Stat.* **4** 1942–1975. MR2829942 <https://doi.org/10.1214/10-AOAS351>
- 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>
- BERROCAL, V. J., GUAN, Y., MUYSKENS, A., WANG, H., REICH, B. J., MULHOLLAND, J. A. and CHANG, H. H. (2019). A comparison of statistical and machine learning methods for creating national daily maps of ambient PM_{2.5} concentration. *Atmos. Environ.* <https://doi.org/10.1016/j.atmosenv.2019.117130>

- BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. [MR0373208](#)
- BESAG, J., GREEN, P., HIGDON, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statist. Sci.* **10** 3–66. With comments and a reply by the authors. [MR1349818](#)
- 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>
- GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. [MR3235677](#)
- GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics, 4* (Peñíscola, 1991) 169–193. Oxford Univ. Press, New York. [MR1380276](#)
- GUAN, Y., REICH, B. J., MULHOLLAND, J. A. and CHANG, H. H. (2019). Multivariate spectral downscaling for PM_{2.5} species. arXiv preprint [arXiv:1909.03816](#).
- GURUNG, A., LEVY, J. I. and BELL, M. L. (2017). Modeling the intraurban variation in nitrogen dioxide in urban areas in Kathmandu Valley, Nepal. *Environ. Res.* **155** 42–48.
- Hsu, C.-Y., Wu, J.-Y., Chen, Y.-C., Chen, N.-T., Chen, M.-J., Pan, W.-C., Lung, S.-C. C., Guo, Y. L. and Wu, C.-D. (2019). Asian culturally specific predictors in a large-scale land use regression model to predict spatial-temporal variability of ozone concentration. *Int. J. Environ. Res. Public Health* **16** 1300.
- JIN, L., BERMAN, J. D., WARREN, J. L., LEVY, J. I., THURSTON, G., ZHANG, Y., XU, X., WANG, S., ZHANG, Y. et al. (2019). A land use regression model of nitrogen dioxide and fine particulate matter in a complex urban core in Lanzhou, China. *Environ. Res.* **177** 108597.
- MCMILLAN, N. J., HOLLAND, D. M., MORARA, M. and FENG, J. (2010). Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics* **21** 48–65. [MR2842223](#) <https://doi.org/10.1002/env.984>
- NOLTE, C. G., SPERO, T. L., BOWDEN, J. H., MALLARD, M. S. and DOLWICK, P. D. (2018). The potential effects of climate change on air quality across the conterminous US at 2030 under three representative concentration pathways. *Atmos. Chem. Phys.* **18** 15471.
- PACIOREK, C. J. (2012). Combining spatial information sources while accounting for systematic errors in proxies. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 429–451. [MR2914520](#) <https://doi.org/10.1111/j.1467-9876.2011.01035.x>
- REICH, B. J., CHANG, H. H. and FOLEY, K. M. (2014). A spectral method for spatial downscaling. *Biometrics* **70** 932–942. [MR3295754](#) <https://doi.org/10.1111/biom.12196>
- TANNER, M. A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd ed. *Springer Series in Statistics*. Springer, New York. [MR1396311](#) <https://doi.org/10.1007/978-1-4612-4024-2>
- US EPA (2019). CMAQ models. <https://www.epa.gov/cmaq/cmaq-models-0>.
- WACKERNAGEL, H. (2013). *Multivariate Geostatistics: An Introduction with Applications*. Springer Science & Business Media.
- WARREN, J. L., KONG, W., LUBEN, T. J. and CHANG, H. H. (2020). Critical window variable selection: Estimating the impact of air pollution on very preterm birth. *Biostatistics* **21** 790–806. <https://doi.org/10.1093/biostatistics/kxz006>
- WARREN, J. L., MIRANDA, M. L., TOOTOO, J. L., OSGOOD, C. E. and BELL, M. L. (2021). Supplement to “Spatial distributed lag data fusion for estimating ambient air pollution.” <https://doi.org/10.1214/20-AOAS1399SUPPA>, <https://doi.org/10.1214/20-AOAS1399SUPPB>
- 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](#)
- XU, H., BECHLE, M. J., WANG, M., SZPIRO, A. A., VEDAL, S., BAI, Y. and MARSHALL, J. D. (2019). National PM_{2.5} and NO₂ exposure models for China based on land use regression, satellite measurements, and universal kriging. *Sci. Total Environ.* **655** 423–433.
- YU, Y., YAO, S., DONG, H., WANG, L., WANG, C., Ji, X., Ji, M., YAO, X. and ZHANG, Z. (2019). Association between short-term exposure to particulate matter air pollution and cause-specific mortality in Changzhou, China. *Environ. Res.* **170** 7–15.

ESTIMATION OF CELL LINEAGE TREES BY MAXIMUM-LIKELIHOOD PHYLOGENETICS

BY JEAN FENG¹, WILLIAM S. DEWITT III², AARON MCKENNA³, NOAH SIMON^{4,*},
 AMY D. WILLIS^{4,†} AND FREDERICK A. MATSEN IV⁵

¹Department of Epidemiology and Biostatistics, University of California, San Francisco, jean.feng@ucsf.edu

²Department of Genome Sciences, University of Washington, wedewitt@uw.edu

³Department of Molecular and Systems Biology, Dartmouth College, aaron.mckenna@dartmouth.edu

⁴Department of Biostatistics, University of Washington, * nrsimon@uw.edu; † adwillis@uw.edu

⁵Computational Biology Program, Fred Hutchinson Cancer Research Center, matsen@fredhutch.org

CRISPR technology has enabled cell lineage tracing for complex multicellular organisms through insertion-deletion mutations of synthetic genomic barcodes during organismal development. To reconstruct the cell lineage tree from the mutated barcodes, current approaches apply general-purpose computational tools that are agnostic to the mutation process and are unable to take full advantage of the data's structure. We propose a statistical model for the CRISPR mutation process and develop a procedure to estimate the resulting tree topology, branch lengths and mutation parameters by iteratively applying penalized maximum likelihood estimation. By assuming the barcode evolves according to a molecular clock, our method infers relative ordering across parallel lineages, whereas existing techniques only infer ordering for nodes along the same lineage. When analyzing transgenic zebrafish data from (*Science* **353** (2016) aaf7907), we find that our method recapitulates known aspects of zebrafish development and the results are consistent across samples.

REFERENCES

- ADAMS, D. C. and COLLYER, M. L. (2018). Multivariate phylogenetic comparative methods: Evaluations, comparisons, and recommendations. *Syst. Biol.* **67** 14–31. <https://doi.org/10.1093/sysbio/syx055>
- ALEMANY, A., FLORESCU, M., BARON, C. S., PETERSON-MADURO, J. and VAN OUDENAARDEN, A. (2018). Whole-organism clone tracing using single-cell sequencing. *Nature* **556** 108–112. <https://doi.org/10.1038/nature25969>
- BILLERA, L. J., HOLMES, S. P. and VOGTMANN, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** 733–767. MR1867931 <https://doi.org/10.1006/aama.2001.0759>
- CAMIN, J. H. and SOKAL, R. R. (1965). A method for deducing branching sequences in phylogeny. *Evolution* **19** 311–326.
- DAVYDOV, I. I., ROBINSON-RECHAVI, M. and SALAMIN, N. (2017). State aggregation for fast likelihood computations in molecular evolution. *Bioinformatics* **33** 354–362. <https://doi.org/10.1093/bioinformatics/btw632>
- DINH, V., HO, L. S. T., SUCHARD, M. A. and MATSEN, F. A. IV (2018). Consistency and convergence rate of phylogenetic inference via regularization. *Ann. Statist.* **46** 1481–1512. MR3819107 <https://doi.org/10.1214/17-AOS1592>
- FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17** 368–376.
- FELSENSTEIN, J. (2004). *Inferring Phylogenies* **2**. Sinauer, Sunderland, MA.
- FENG, J., DEWITT, W. S., MCKENNA, A., SIMON, N., WILLIS, A. and MATSEN, F. A. (2021). Supplement to “Estimation of cell lineage trees by maximum-likelihood phylogenetics.” <https://doi.org/10.1214/20-AOAS1400SUPPA>, <https://doi.org/10.1214/20-AOAS1400SUPPB>
- GANGULY, A., PETROV, T. and KOEPL, H. (2014). Markov chain aggregation and its applications to combinatorial reaction networks. *J. Math. Biol.* **69** 767–797. MR3248074 <https://doi.org/10.1007/s00285-013-0738-7>

Key words and phrases. Phylogenetics, tree reconstruction, cell lineage tracing, maximum likelihood estimator, regularization.

- GOOLSBY, E. W. (2016). Likelihood-based parameter estimation for high-dimensional phylogenetic comparative models: Overcoming the limitations of “distance-based” methods. *Syst. Biol.* **65** 852–870.
- 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) <https://doi.org/10.1007/978-0-387-84858-7>
- HILLSTON, J. (1995). Compositional Markovian modelling using a process algebra. In *Computations with Markov Chains* 177–196. Springer, US.
- JULIEN, C., LEANDRO, A. and HÉLÈNE, M. (2018). A penalized likelihood framework for high-dimensional phylogenetic comparative methods and an application to new-world monkeys brain evolution. *Syst. Biol.*
- KALHOR, R., MALI, P. and CHURCH, G. M. (2017). Rapidly evolving homing CRISPR barcodes. *Nat. Methods* **14** 195–200. <https://doi.org/10.1038/nmeth.4108>
- KALHOR, R., KALHOR, K., MEJIA, L., LEEPER, K., GRAVELINE, A., MALI, P. and CHURCH, G. M. (2018). Developmental barcoding of whole mouse via homing CRISPR. *Science* **361**. <https://doi.org/10.1126/science.aaf9804>
- KEEGAN, B. R., MEYER, D. and YELON, D. (2004). Organization of cardiac chamber progenitors in the zebrafish blastula. *Development* **131** 3081–3091.
- KEMENY, J. G. and SNELL, J. L. (1976). *Finite Markov Chains. Undergraduate Texts in Mathematics*. Springer, New York. [MR0410929](https://doi.org/10.1007/978-1-4757-3814-1)
- KIM, J. and SANDERSON, M. J. (2008). Penalized likelihood phylogenetic inference: Bridging the parsimony-likelihood gap. *Syst. Biol.* **57** 665–674.
- MCKENNA, A. and GAGNON, J. A. (2019). Recording development with single cell dynamic lineage tracing. *Development* **146**. <https://doi.org/10.1242/dev.169730>
- MCKENNA, A., FINDLAY, G. M., GAGNON, J. A., HORWITZ, M. S., SCHIER, A. F. and SHENDURE, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353** aaf7907. <https://doi.org/10.1126/science.aaf7907>
- MOODY, S. A. (1998). *Cell Lineage and Fate Determination*. Elsevier, Amsterdam.
- SAITOU, N. and NEI, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4** 406–425.
- SALVADOR-MARTÍNEZ, I., GRILLO, M., AVEROF, M. and TELFORD, M. J. (2018). Is it possible to reconstruct an accurate cell lineage using CRISPR recorders?
- SANDERSON, M. J. (2002). Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Mol. Biol. Evol.* **19** 101–109.
- SCHMIDT, S. T., ZIMMERMAN, S. M., WANG, J., KIM, S. K. and QUAKE, S. R. (2017). Quantitative analysis of synthetic cell lineage tracing using nuclease barcoding. *ACS Synth. Biol.* **6** 936–942.
- SHADRIN, A. M. and OZERNYUK, N. D. (2002). Development of the gill system in early ontogenesis of the zebrafish and ninespine stickleback. *Russ. J. Dev. Biol.* **33** 91–96.
- SOLNICA-KREZEL, L. (2005). Conserved patterns of cell movements during vertebrate gastrulation. *Curr. Biol.* **15** R213–R228. <https://doi.org/10.1016/j.cub.2005.03.016>
- SPANJAARD, B., HU, B., MITIC, N., OLIVARES-CHAUVEY, P., JANJUHA, S., NINOV, N. and JUNKER, J. P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36** 469–473.
- WOODWORTH, M. B., GIRSKIS, K. M. and WALSH, C. A. (2017). Building a lineage from single cells: Genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* **18** 230–244. <https://doi.org/10.1038/nrg.2016.159>
- WU, Y.-C., RASMUSSEN, M. D., BANSAL, M. S. and KELLIS, M. (2013). TreeFix: Statistically informed gene tree error correction using species trees. *Syst. Biol.* **62** 110–120.
- YANG, Z. (2014). *Molecular Evolution: A Statistical Approach*. Oxford Univ. Press, London.

BETS: THE DANGERS OF SELECTION BIAS IN EARLY ANALYSES OF THE CORONAVIRUS DISEASE (COVID-19) PANDEMIC

BY QINGYUAN ZHAO^{1,*}, NIANQIAO JU², SERGIO BACALLADO^{1,†} AND
RAJEN D. SHAH^{1,‡}

¹Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge,

*qyzhao@statslab.cam.ac.uk; †sb2116@cam.ac.uk; ‡R.Shah@statslab.cam.ac.uk

²Department of Statistics, Harvard University, nju@g.harvard.edu

The coronavirus disease 2019 (COVID-19) has quickly grown from a regional outbreak in Wuhan, China, to a global pandemic. Early estimates of the epidemic growth and incubation period of COVID-19 may have been biased due to sample selection. Using detailed case reports from 14 locations in and outside mainland China, we obtained 378 Wuhan-exported cases who left Wuhan before an abrupt travel quarantine. We developed a generative model we call BETS for four key epidemiological events—Beginning of exposure, End of exposure, time of Transmission, and time of Symptom onset (BETS)—and derived explicit formulas to correct for the sample selection. We gave a detailed illustration of why some early and highly influential analyses of the COVID-19 pandemic were severely biased. All our analyses, regardless of which subsample and model were being used, point to an epidemic doubling time of *two* to 2.5 days during the early outbreak in Wuhan. A Bayesian nonparametric analysis further suggests that about 5% of the symptomatic cases may not develop symptoms within 14 days of infection and that men may be much more likely than women to develop symptoms within two days of infection.

REFERENCES

- BACKER, J. A., KLINKENBERG, D. and WALLINGA, J. (2020). Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. *Euro Surveill.* **25**. <https://doi.org/10.2807/1560-7917.ES.2020.25.5.2000062>
- CENTERS FOR DISEASE CONTROL AND PREVENTION. Interim clinical guidance for management of patients with confirmed coronavirus disease (COVID-19). Available at <https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html>. Retrieved: April 15, 2020.
- DILLON, J. V., LANGMORE, I., TRAN, D., BREVDO, E., VASUDEVAN, S., MOORE, D., PATTON, B., ALEMI, A., HOFFMAN, M. et al. (2017). Tensorflow distributions. Preprint. Available at [arXiv:1711.10604](https://arxiv.org/abs/1711.10604).
- DORIGATTI, I., OKELL, L., CORI, A., IMAI, N., BAGUELIN, M., BHATIA, S., BOONYASIRI, A., CUCUNUBÁ, Z., CUOMO-DANNENBURG, G. et al. (2020). Report 4—Severity of 2019 novel coronavirus (nCoV). Technical report, Imperial College London, MRC Centre for Global Infectious Disease Analysis.
- DU, Z., XU, X., WU, Y., WANG, L., COWLING, B. J. and MEYERS, L. A. (2020). The serial interval of COVID-19 from publicly reported confirmed cases. *Emerg. Infect. Dis.* **26**.
- EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability **57**. CRC Press, New York. MR1270903 <https://doi.org/10.1007/978-1-4899-4541-9>
- GELMAN, A. (2020). Concerns with that Stanford study of coronavirus prevalence. Available at <https://statmodeling.stat.columbia.edu/2020/04/19/fatal-flaws-in-stanford-study-of-coronavirus-prevalence/>.
- GELMAN, A., RUBIN, D. B. et al. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
- GLOBAL HEALTH 5050. COVID-19 sex-disaggregated data tracker. Available at <https://globalhealth5050.org/covid19/>. Retrieved: April 28, 2020.
- HÁJEK, J., ŠIDÁK, Z. and SEN, P. K. (1999). *Theory of Rank Tests*, 2nd ed. Probability and Mathematical Statistics. Academic Press, San Diego, CA. MR1680991

Key words and phrases. Epidemiology, infectious disease, Bayesian nonparametrics, selection bias, epidemic growth, incubation period.

- HETHCOTE, H. W. (2000). The mathematics of infectious diseases. *SIAM Rev.* **42** 599–653. [MR1814049 https://doi.org/10.1137/S0036144500371907](https://doi.org/10.1137/S0036144500371907)
- IMAI, N., DORIGATTI, I., CORI, A., RILEY, S. and FERGUSON, N. M. (2020). Report 1: Estimating the potential total number of novel Coronavirus cases in Wuhan City, China. Technical report, Imperial College London, MRC Centre for Global Infectious Disease Analysis.
- JOHNS HOPKINS UNIVERSITY. COVID-19 dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). Available at <https://coronavirus.jhu.edu/map.html>. Retrieved: April 15, 2020.
- KERMACK, W. O. and MCKENDRICK, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A* **115** 700–721.
- LAUER, S. A., GRANTZ, K. H., BI, Q., JONES, F. K., ZHENG, Q., MEREDITH, H. R., AZMAN, A. S., REICH, N. G. and LESSLER, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann. Intern. Med.* **172** 577–582.
- LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. Clarendon Press, New York. [MR1419991 https://doi.org/10.1093/oxfordjbs.17.1.1](https://doi.org/10.1093/oxfordjbs.17.1.1)
- LI, Q., GUAN, X., WU, P., WANG, X., ZHOU, L., TONG, Y., REN, R., LEUNG, K. S., LAU, E. H., WONG, J. Y. et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382** 1199–1207.
- LINTON, N. M., KOBAYASHI, T., YANG, Y., HAYASHI, K., AKHMETZHANOV, A. R., JUNG, S.-M., YUAN, B., KINOSHITA, R. and NISHIURA, H. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *J. Clin. Med.* **9** E538.
- LIPSITCH, M. (2020). Estimating case fatality rates of COVID-19. *Lancet, Infect. Dis.*
- LIPSITCH, M., JOSHI, K. and COBEY, S. E. (2020). Comment on “Association of public health interventions with the epidemiology of the COVID-19 outbreak in Wuhan, China” by Pan et al. Available at https://github.com/keyajoshi/Pan_response.
- NIEHUS, R., MARTINEZ DE SALAZAR, P., TAYLOR, A. R. and LIPSITCH, M. (2020). Using observational data to quantify bias of traveller-derived COVID-19 prevalence estimates in Wuhan, China. *Lancet* **20** 803–808. [https://doi.org/10.1016/S1473-3099\(20\)30229-2](https://doi.org/10.1016/S1473-3099(20)30229-2)
- NISHIURA, H., LINTON, N. M. and AKHMETZHANOV, A. R. (2020). Serial interval of novel coronavirus (COVID-19) infections. *Int. J. Infect. Dis.* **93** 284–286. <https://doi.org/10.1016/j.ijid.2020.02.060>
- READ, J. M., BRIDGEN, J. R., CUMMINGS, D. A., HO, A. and JEWELL, C. P. (2020). Novel coronavirus 2019-nCoV: Early estimation of epidemiological parameters and epidemic predictions. *medRxiv*. <https://doi.org/10.1101/2020.01.23.20018549>
- REICH, N. G., LESSLER, J., CUMMINGS, D. A. T. and BROOKMEYER, R. (2009). Estimating incubation period distributions with coarse data. *Stat. Med.* **28** 2769–2784. [MR2750164 https://doi.org/10.1002/sim.3659](https://doi.org/10.1002/sim.3659)
- SANCHE, S., LIN, Y., XU, C., ROMERO-SEVERSON, E., HENGARTNER, N. and KE, R. (2020). High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerg. Infect. Dis.* **26**.
- UK GOVERNMENT (2020a). Boris Johnson coronavirus speech transcript: UK PM tells UK to avoid non-essential travel & contact. Available at <https://www.rev.com/blog/transcripts/boris-johnson-coronavirus-speech-transcript-uk-pm-tells-uk-to-avoid-non-essential-travel-contact>. Retrieved: April 15, 2020.
- UK GOVERNMENT (2020b). Michael Gove: Rate of coronavirus infection in UK doubling every three to four days—Video. Available at <https://www.theguardian.com/world/video/2020/mar/27/michael-gove-rate-of-coronavirus-infection-in-uk-doubling-every-three-to-four-days-video>. Retrieved: April 15, 2020.
- WALLINGA, J. and LIPSITCH, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. Lond., B Biol. Sci.* **274** 599–604.
- WILLEM, L., VERELST, F., BILCKE, J., HENS, N. and BEUTELS, P. (2017). Lessons from a decade of individual-based models for infectious disease transmission: A systematic review (2006–2015). *BMC Infect. Dis.* **17** 612. <https://doi.org/10.1186/s12879-017-2699-8>
- WONG, J. C. and BERGSTROM, C. (2020). ‘There is no absolute truth’: An infectious disease expert on Covid-19, misinformation and ‘bullshit’. Available at <https://www.theguardian.com/world/2020/apr/28/there-is-no-absolute-truth-an-infectious-disease-expert-on-covid-19-misinformation-and-bullshit>. Retrieved: May 3, 2020.
- WORLD HEALTH ORGANIZATION (2020). WHO statement regarding cluster of pneumonia cases in Wuhan, China. Available at <https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china>. Retrieved: April 15, 2020.
- WU, J. T., LEUNG, K. and LEUNG, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. *Lancet* **395** 689–697. [https://doi.org/10.1016/S0140-6736\(20\)30260-9](https://doi.org/10.1016/S0140-6736(20)30260-9)

- ZHAO, Q., CHEN, Y. and SMALL, D. S. (2020). Analysis of the epidemic growth of the early 2019-nCoV outbreak using internationally confirmed cases. *medRxiv*. <https://doi.org/10.1101/2020.02.06.20020941>
- ZHAO, Q., JU, N., BACALLADO, S. and SHAH, R. D. (2021). Supplement to “BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic.” <https://doi.org/10.1214/20-AOAS1401SUPPA>, <https://doi.org/10.1214/20-AOAS1401SUPPB>

BAYICE: A BAYESIAN HIERARCHICAL MODEL FOR SEMIREFERENCE-BASED DECONVOLUTION OF BULK TRANSCRIPTOMIC DATA

BY AN-SHUN TAI^{1,*}, GEORGE C. TSENG² AND WEN-PING HSIEH^{1,†}

¹*Institute of Statistics, National Tsing Hua University, *anshantai@nctu.edu.tw; †wphsieh@stat.nthu.edu.tw*

²*Department of Biostatistics, University of Pittsburgh, ctseng@pitt.edu*

Gene expression deconvolution is a powerful tool for exploring the microenvironment of complex tissues comprised of multiple cell groups using transcriptomic data. Characterizing cell activities for a particular condition has been regarded as a primary mission against diseases. For example, cancer immunology aims to clarify the role of the immune system in the progression and development of cancer through analyzing the immune cell components of tumors. To that end, many deconvolution methods have been proposed for inferring cell subpopulations within tissues. Nevertheless, two problems limit the practicality of current approaches. First, most approaches use external purified data to preselect cell type-specific genes that contribute to deconvolution. However, some types of cells cannot be found in purified profiles, and the genes specifically over- or under-expressed in them cannot be identified. This is particularly a problem in cancer studies. Hence, a preselection strategy that is independent from deconvolution is inappropriate. The second problem is that existing approaches do not recover the expression profiles of unknown cells present in bulk tissues when the reference set of purified cell-specific profiles is incomplete which results in biased estimation of unknown cell proportions. Furthermore, it causes the shift-invariant property of deconvolution to fail which then affects the estimation performance. To address these two problems, we propose a novel semireference-based deconvolution approach, BayICE which employs hierarchical Bayesian modeling with stochastic search variable selection. We develop a comprehensive Markov chain Monte Carlo procedure through Gibbs sampling to estimate proportions, expression profiles and signature genes for a set of known reference cell types as well as an unknown cell type. Simulation and validation studies illustrate that BayICE outperforms existing semireference-based deconvolution approaches in estimating cell proportions. We further show that BayICE is applicable to single-cell RNA-seq data. Subsequently, we demonstrate an application of BayICE in the RNA sequencing of patients with non-small cell lung cancer. The model is implemented in the R package “BayICE,” and the algorithm is available for download.

REFERENCES

- ABBAS, A. R., WOLSLEGEL, K., SESHASAYEE, D., MODRUSAN, Z. and CLARK, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE* **4** e6098. <https://doi.org/10.1371/journal.pone.0006098>
- ANGHEL, C. V., QUON, G., HAIDER, S., NGUYEN, F., DESHWAR, A. G., MORRIS, Q. D. and BOUTROS, P. C. (2015). ISOPureR: An R implementation of a computational purification algorithm of mixed tumour profiles. *BMC Bioinform.* **16** 156.
- BANAT, G., TRETYN, A., PULLAMSETTI, S. S., WILHELM, J., WEIGERT, A., OLESCH, C., EBEL, K., STIEWE, T., GRIMMINGER, F. et al. (2015). Immune and inflammatory cell composition of human lung cancer stroma. *PLoS ONE* **10** e0139073. <https://doi.org/10.1371/journal.pone.0139073>

- BECHT, E., GIRALDO, N. A., LACROIX, L., BUTTARD, B., ELAROUCI, N., PETITPREZ, F., SELVES, J., LAURENT-PUIG, P., SAUTÈS-FRIDMAN, C. et al. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17** 218.
- CHAROENTONG, P., FINOTELLO, F., ANGELOVA, M., MAYER, C., EFREMOVA, M., RIEDER, D., HACKL, H. and TRAJANOSKI, Z. (2017). Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* **18** 248–262. <https://doi.org/10.1016/j.celrep.2016.12.019>
- CHEN, Z., HUANG, A., SUN, J., JIANG, T., QIN, F. X.-F. and WU, A. (2017). Inference of immune cell composition on the expression profiles of mouse tissue. *Sci. Rep.* **7** 40508.
- CHEN, Z., QUAN, L., HUANG, A., ZHAO, Q., YUAN, Y., YUAN, X., SHEN, Q., SHANG, J., BEN, Y. et al. (2018). seq-ImmuCC: Cell-centric view of tissue transcriptome measuring cellular compositions of immune microenvironment from mouse RNA-Seq data. *Front. Immunol.* **9** 1286. <https://doi.org/10.3389/fimmu.2018.01286>
- DIEU-NOSJEAN, M.-C., GOC, J., GIRALDO, N. A., SAUTÈS-FRIDMAN, C. and FRIDMAN, W. H. (2014). Tertiary lymphoid structures in cancer and beyond. *Trends Immunol.* **35** 571–580.
- FRIDMAN, W. H., PAGÈS, F., SAUTÈS-FRIDMAN, C. and GALON, J. (2012). The immune contexture in human tumours: Impact on clinical outcome. *Nat. Rev. Cancer* **12** nrc3245.
- FRIDMAN, W. H., ZITVOGEL, L., SAUTÈS-FRIDMAN, C. and KROEMER, G. (2017). The immune contexture in cancer prognosis and treatment. *Nat. Rev. Clin. Oncol.* **14** 717.
- GAUJOUX, R. and SEOIGHE, C. (2012). Semi-supervised nonnegative matrix factorization for gene expression deconvolution: A case study. *Infect. Genet. Evol.* **12** 913–921.
- GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.
- GONG, T., HARTMANN, N., KOHANE, I. S., BRINKMANN, V., STAEDTLER, F., LETZKUS, M., BONGIOVANNI, S. and SZUSTAKOWSKI, J. D. (2011). Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS ONE* **6** e27156. <https://doi.org/10.1371/journal.pone.0027156>
- HU, P., ZHANG, W., XIN, H. and DENG, G. (2016). Single cell isolation and analysis. *Front. Cell Dev. Biol.* **4** 116. <https://doi.org/10.3389/fcell.2016.00116>
- HUNT, G. J., FREYTAG, S., BAHLO, M. and GAGNON-BARTSCH, J. A. (2018). dtangle: Accurate and robust cell type deconvolution. *Bioinformatics* **35** 2093–2099.
- ISHWARAN, H., KOGALUR, U. B. and RAO, J. S. (2010). spikeslab: Prediction and variable selection using spike and slab regression. *R J.* **2**.
- ISHWARAN, H. and RAO, J. S. (2005a). Spike and slab gene selection for multigroup microarray data. *J. Amer. Statist. Assoc.* **100** 764–780. [MR2201009 https://doi.org/10.1198/016214505000000051](https://doi.org/10.1198/016214505000000051)
- ISHWARAN, H. and RAO, J. S. (2005b). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. [MR2163158 https://doi.org/10.1214/009053604000001147](https://doi.org/10.1214/009053604000001147)
- KOPF, M., SCHNEIDER, C. and NOBS, S. P. (2015). The development and function of lung-resident macrophages and dendritic cells. *Nat. Immunol.* **16** 36–44. <https://doi.org/10.1038/ni.3052>
- LI, B., SEVERSON, E., PIGNON, J.-C., ZHAO, H., LI, T., NOVAK, J., JIANG, P., SHEN, H., ASTER, J. C. et al. (2016). Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biol.* **17** 174. <https://doi.org/10.1186/s13059-016-1028-7>
- MEZHEYEUSKI, A., BERGLAND, C. H., BACKMAN, M., DJUREINOVIC, D., SJÖBLOM, T., BRUUN, J. and MICKE, P. (2018). Multispectral imaging for quantitative and compartment-specific immune infiltrates reveals distinct immune profiles that classify lung cancer patients. *J. Pathol.* **244** 421–431. <https://doi.org/10.1002/path.5026>
- NEWMAN, A. M., LIU, C. L., GREEN, M. R., GENTLES, A. J., FENG, W., XU, Y., HOANG, C. D., DIEHN, M. and ALIZADEH, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12** 453–457. <https://doi.org/10.1038/nmeth.3337>
- NEWMAN, A. M., STEEN, C. B., LIU, C. L., GENTLES, A. J., CHAUDHURI, A. A., SCHERER, F., KHODADOUST, M. S., ESFAHANI, M. S., LUCA, B. A. et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37** 773–782. <https://doi.org/10.1038/s41587-019-0114-2>
- OGUNDIJO, O. E. and WANG, X. (2017). A sequential Monte Carlo approach to gene expression deconvolution. *PLoS ONE* **12** e0186167.
- PABST, C., BERGERON, A., LAVALLÉE, V.-P., YEH, J., GENDRON, P., NORDDAHL, G. L., KROSL, J., BOIVIN, I., DENEALT, E. et al. (2016). GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. *Blood*:blood-2015-2011-683649.
- PRASSAS, I., CHRYSOIA, C. C., MAKAWITA, S. and DIAMANDIS, E. P. (2012). Bioinformatic identification of proteins with tissue-specific expression for biomarker discovery. *BMC Med.* **10** 39.

- RACLE, J., DE JONGE, K., BAUMGAERTNER, P., SPEISER, D. E. and GFELLER, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**. <https://doi.org/10.7554/eLife.26476>
- RICHARDSON, G. M., LANNIGAN, J. and MACARA, I. G. (2015). Does FACS perturb gene expression? *Cytometry, Part A* **87** 166–175.
- RUFFINI, E., ASIOLI, S., FILOSSO, P. L., LYBERIS, P., BRUNA, M. C., MACRÌ, L., DANIELE, L. and OLIARO, A. (2009). Clinical significance of tumor-infiltrating lymphocytes in lung neoplasms. *Ann. Thorac. Surg.* **87** 365–372.
- SEO, J.-S., KIM, A., SHIN, J.-Y. and KIM, Y. T. (2018). Comprehensive analysis of the tumor immune micro-environment in non-small cell lung cancer for efficacy of checkpoint inhibitor. *Sci. Rep.* **8** 14576.
- SHEN-ORR, S. S., TIBSHIRANI, R., KHATRI, P., BODIAN, D. L., STAEDTLER, F., PERRY, N. M., HASTIE, T., SARWAL, M. M., DAVIS, M. M. et al. (2010). Cell type-specific gene expression differences in complex tissues. *Nat. Methods* **7** 287–289. <https://doi.org/10.1038/nmeth.1439>
- SUN, X., SUN, S. and YANG, S. (2019). An efficient and flexible method for deconvoluting bulk RNA-seq data with single-cell RNA-seq data. *Cells* **8** 1161.
- TAI, A.-S., TSENG, G. C. and HSIEH, W.-P. (2021). Supplement to “BayICE: A Bayesian hierarchical model for semireference-based deconvolution of bulk transcriptomic data.” <https://doi.org/10.1214/20-AOAS1376SUPP>
- TIROSH, I., IZAR, B., PRAKADAN, S. M., WADSWORTH, M. H., TREACY, D., TROMBETTA, J. J., ROTEM, A., RODMAN, C., LIAN, C. et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352** 189–196.
- TRAN, H. T. N., ANG, K. S., CHEVRIER, M., ZHANG, X., LEE, N. Y. S., GOH, M. and CHEN, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21** 1–32.
- VAN DEN BRINK, S. C., SAGE, F., VÉRTESY, Á., SPANJAARD, B., PETERSON-MADURO, J., BARON, C. S., ROBIN, C. and VAN OUDENAARDEN, A. (2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14** 935–936. <https://doi.org/10.1038/nmeth.4437>
- WAGNER, G. P., KIN, K. and LYNCH, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131** 281–285.
- WANG, X., PARK, J., SUSZTAK, K., ZHANG, N. R. and LI, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10** 1–9.
- WHITTEMORE, A. S. (2007). A Bayesian false discovery rate for multiple testing. *J. Appl. Stat.* **34** 1–9. <https://doi.org/10.1080/02664760600994745>
- XIE, F., ZHOU, M. and XU, Y. (2018). BayCount: A Bayesian decomposition method for inferring tumor heterogeneity using RNA-Seq counts. *Ann. Appl. Stat.* **12** 1605–1627. <https://doi.org/10.1214/17-AOAS1123>
- YANG, Y. H., DUDOIT, S., LUU, P., LIN, D. M., PENG, V., NGAI, J. and SPEED, T. P. (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30** e15–e15.
- YOSHIHARA, K., SHAHMORADGOLI, M., MARTÍNEZ, E., VEGESNA, R., KIM, H., TORRES-GARCIA, W., TREVIÑO, V., SHEN, H., LAIRD, P. W. et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4** 2612. <https://doi.org/10.1038/ncomms3612>
- ZHONG, Y., WAN, Y.-W., PANG, K., CHOW, L. M. and LIU, Z. (2013). Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinform.* **14** 89.

A MULTIPLE IMPUTATION PROCEDURE FOR RECORD LINKAGE AND CAUSAL INFERENCE TO ESTIMATE THE EFFECTS OF HOME-DELIVERED MEALS

BY MINGYANG SHAN^{1,*}, KALI S. THOMAS² AND ROEE GUTMAN^{1,†}

¹Department of Biostatistics, Brown University School of Public Health, *mingyang_shan@brown.edu;
[†]roee_gutman@brown.edu

²Department of Health Services, Policy and Practice, Brown University School of Public Health, kali_thomas@brown.edu

Causal analysis of observational studies requires data that comprise a set of covariates, a treatment assignment indicator and the observed outcomes. However, data confidentiality restrictions or the nature of data collection may distribute these variables across two or more datasets. In the absence of unique identifiers to link records across files, probabilistic record linkage algorithms can be leveraged to merge the datasets. Current applications of record linkage are concerned with estimation of associations between variables that are exclusive to one file and not causal relationships. We propose a Bayesian framework for record linkage and causal inference where one file comprises all the covariate and observed outcome information, and the second file consists of a list of all individuals who receive the active treatment. Under certain ignorability assumptions, the procedure properly propagates the error in the record linkage process, resulting in valid statistical inferences. To estimate the causal effects, we devise a two-stage procedure. The first stage of the procedure performs Bayesian record linkage to multiply-impute the treatment assignment for all individuals in the first file, while adjustments for covariates' imbalance and imputation of missing potential outcomes are performed in the second stage. This procedure is used to evaluate the effect of Meals on Wheels services on mortality and healthcare utilization among homebound older adults in Rhode Island. In addition, an interpretable sensitivity analysis is developed to assess potential violations of the ignorability assumptions.

REFERENCES

- ABADIE, A. and IMBENS, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *J. Bus. Econom. Statist.* **29** 1–11. [MR2789386 https://doi.org/10.1198/jbes.2009.07333](https://doi.org/10.1198/jbes.2009.07333)
- CAMPBELL, K. M., DECK, D. and KRUPSKI, A. (2008). Record linkage software in the public domain: A comparison of Link Plus, The Link King, and a 'basic' deterministic algorithm. *Health Inform. J.* **14** 5–15.
- CAMPBELL, A. D., GODFRYD, A., BUYS, D. R. and LOCHER, J. L. (2015). Does participation in home-delivered meals programs improve outcomes for older adults? Results of a systematic review. *J. Nutr. Gerontol. Geriatr.* **34** 124–167. <https://doi.org/10.1080/21551197.2015.1038463>
- CHAMBERS, R., CHIPPERFIELD, J., DAVIS, W. and KOVACEVIC, M. (2009). Inference based on estimating equations and probability-linked data. Centre for Statistical and Survey Methodology, University of Wollongong Working Paper Series 18-09.
- COPAS, J. B. and HILTON, F. J. (1990). Record linkage: Statistical models for matching computer records. *J. Roy. Statist. Soc. Ser. A* **153** 287–320.
- FELLEGI, I. P. and SUNTER, A. B. (1969). A theory for record linkage. *J. Amer. Statist. Assoc.* **64** 1183–1210.
- FORTINI, M., LISEO, B. and NUCCITELLI, A. (2001). On Bayesian record linkage **4** 185–198.
- 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)
- FRANGAKIS, C. E., RUBIN, D. B., AN, M.-W. and MACKENZIE, E. (2007). Principal stratification designs to estimate input data missing due to death. *Biometrics* **63** 641–649. [MR2395697 https://doi.org/10.1111/j.1541-0420.2007.00847_1.x](https://doi.org/10.1111/j.1541-0420.2007.00847_1.x)

Key words and phrases. Record linkage, missing data, causal inference, multiple imputation, Bayesian data analysis.

- GOMATAM, S., CARTER, R., ARIET, M. and MITCHELL, G. (2002). An empirical comparison of record linkage procedures. *Stat. Med.* **21** 1485–1496.
- GRIFFIN, J. and BROWN, P. (2017). Hierarchical shrinkage priors for regression models. *Bayesian Anal.* **12** 135–159. [MR3597570](#) <https://doi.org/10.1214/15-BA990>
- GUTMAN, R., AFENDULIS, C. C. and ZASLAVSKY, A. M. (2013). A Bayesian procedure for file linking to analyze end-of-life medical costs. *J. Amer. Statist. Assoc.* **108** 34–47. [MR3174601](#) <https://doi.org/10.1080/01621459.2012.726889>
- GUTMAN, R. and RUBIN, D. B. (2013). Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes. *Stat. Med.* **32** 1795–1814. [MR3067363](#) <https://doi.org/10.1002/sim.5627>
- GUTMAN, R. and RUBIN, D. B. (2015). Estimation of causal effects of binary treatments in unconfounded studies. *Stat. Med.* **34** 3381–3398. [MR3412639](#) <https://doi.org/10.1002/sim.6532>
- GUTMAN, R. and RUBIN, D. B. (2017). Estimation of causal effects of binary treatments in unconfounded studies with one continuous covariate. *Stat. Methods Med. Res.* **26** 1199–1215. [MR3660989](#) <https://doi.org/10.1177/0962280215570722>
- HARRON, K., GOLDSTEIN, H. and DIBBEN, C. (2015). *Methodological Developments in Data Linkage*. Wiley Series in Probability and Statistics. Wiley, New York.
- HERZOG, T., SCHEUREN, F. and WINKLER, W. (2007). *Data Quality and Record Linkage Techniques*. Springer, New York.
- HOF, M. H. P. and ZWINDERMAN, A. H. (2012). Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Stat. Med.* **31** 4231–4242. [MR3040077](#) <https://doi.org/10.1002/sim.5498>
- IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. [MR3309951](#) <https://doi.org/10.1017/CBO9781139025751>
- JARO, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *J. Amer. Statist. Assoc.* **84** 414–420.
- LAHIRI, P. and LARSEN, M. D. (2005). Regression analysis with linked data. *J. Amer. Statist. Assoc.* **100** 222–230. [MR2156832](#) <https://doi.org/10.1198/016214504000001277>
- LARSEN, M. D. (2005). Hierarchical Bayesian record linkage theory.
- LARSEN, M. D. and RUBIN, D. B. (2001). Iterative automated record linkage using mixture models. *J. Amer. Statist. Assoc.* **96** 32–41. [MR1973781](#) <https://doi.org/10.1198/016214501750332956>
- LEE, J. S., SHANNON, J. and BROWN, A. (2015). Characteristics of older Georgians receiving older Americans act nutrition program services and other home- and community-based services: Findings from the Georgia aging information management system (GA AIMS). *J. Nutr. Gerontol. Geriatr.* **34** 168–188.
- LLOYD, J. L. and WELLMAN, N. S. (2015). Older americans act nutrition programs: A community-based nutrition program helping older adults remain at home. *J. Nutr. Gerontol. Geriatr.* **34** 90–109. [https://doi.org/10.1080/21551197.2015.1031592](#)
- MURRAY, J. S. (2015). Probabilistic record linkage and deduplication after indexing, blocking, and filtering. *J. Priv. Confid.* **7**.
- NETER, J., MAYNES, E. S. and RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response errors. *J. Amer. Statist. Assoc.* **60** 1005–1027. [MR0193729](#)
- NEWCOMBE, H. B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford Univ. Press, Oxford.
- NEWCOMBE, H. B., KENNEDY, J. M., AXFORD, S. J. and JAMES, A. P. (1959). Automatic linkage of vital records. *Science* **130** 954–959.
- NEWMAN, L. M., SAMUEL, M. C., STENGER, M. R., GERBER, T. M., MACOMBER, K., STOVER, J. A. and WISE, W. (2009). Practical considerations for matching STD and HIV surveillance data with data from other sources. *Public Health Rep.* **124** 7–17. [https://doi.org/10.1177/00333549091240S203](#)
- NEYMAN, J. (1923). Sur les applications de la th  rie des probabilit  s aux experiences agricoles: Essay de principe. English translation of excerpts by Dabrowska, D. and Speed, T. (1990). *Statist. Sci.* **5** 465–472.
- ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. [MR0742974](#) <https://doi.org/10.1093/biomet/70.1.41>
- ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.
- RUBIN, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics* **29** 159–183.
- RUBIN, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29** 185–203.
- RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. [MR0472152](#)
- RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–328.

- RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.
- RUBIN, D. B. (1990). Formal mode of statistical inference for causal effects. *J. Statist. Plann. Inference* **25** 279–292.
- RUBIN, D. B. (1996). Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.* **91** 473–489.
- RUBIN, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Stat. Neerl.* **57** 3–18. [MR2055518 https://doi.org/10.1111/1467-9574.00217](https://doi.org/10.1111/1467-9574.00217)
- RUBIN, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with “censoring” due to death. *Statist. Sci.* **21** 299–309. [MR2339125 https://doi.org/10.1214/088342306000000114](https://doi.org/10.1214/088342306000000114)
- RUBIN, D. B. (2008). Statistical inference for causal effects, with emphasis on applications in epidemiology and medical statistics. In *Epidemiology and Medical Statistics. Handbook of Statist.* **27** 28–63. Elsevier/North-Holland, Amsterdam. [MR2500431 https://doi.org/10.1016/S0169-7161\(07\)27002-6](https://doi.org/10.1016/S0169-7161(07)27002-6)
- SADINLE, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *J. Amer. Statist. Assoc.* **112** 600–612. [MR3671755 https://doi.org/10.1080/01621459.2016.1148612](https://doi.org/10.1080/01621459.2016.1148612)
- SADINLE, M. and FIENBERG, S. E. (2013). A generalized Fellegi–Sunter framework for multiple record linkage with application to homicide record systems. *J. Amer. Statist. Assoc.* **108** 385–397. [MR3174628 https://doi.org/10.1080/01621459.2012.757231](https://doi.org/10.1080/01621459.2012.757231)
- SCHUREN, F. and WINKLER, W. (1993). Regression analysis of data files that are computer matched—Part I. *Surv. Methodol.* **19**.
- SCHUREN, F. and WINKLER, W. (1997). Regression analysis of data files that are computer matched—Part II. *Surv. Methodol.* **23** 157–165.
- SHAN, M., THOMAS, K. S. and GUTMAN, R. (2021). Supplement to “A multiple imputation procedure for record linkage and causal inference to estimate the effects of home-delivered meals.” <https://doi.org/10.1214/20-AOAS1397SUPP>
- SHEN, Z. (2000). Nested multiple imputations. Thesis (Ph.D.)—Harvard University. [MR2700720](https://doi.org/10.1080/01621459.2000.1055518)
- STAN DEVELOPMENT TEAM (2018). RStan: The R interface to Stan. R package version 2.17.3.
- STEORTS, R. C. (2015). Entity resolution with empirically motivated priors. *Bayesian Anal.* **10** 849–875. [MR3432242 https://doi.org/10.1214/15-BA965SI](https://doi.org/10.1214/15-BA965SI)
- STEORTS, R. C., HALL, R. and FIENBERG, S. E. (2016). A Bayesian approach to graphical record linkage and deduplication. *J. Amer. Statist. Assoc.* **111** 1660–1672. [MR3601725 https://doi.org/10.1080/01621459.2015.1105807](https://doi.org/10.1080/01621459.2015.1105807)
- STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. [MR2741812 https://doi.org/10.1214/09-STS313](https://doi.org/10.1214/09-STS313)
- TANCREDI, A. and LISEO, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *Ann. Appl. Stat.* **5** 1553–1585. [MR2849786 https://doi.org/10.1214/10-AOAS447](https://doi.org/10.1214/10-AOAS447)
- 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/10.1080/01621459.1987.1047684)
- THOMAS, K. S., AKOBUNDU, U. and DOSA, D. (2015). More than a meal? A randomized control trial comparing the effects of home-delivered meals programs on participants’ feelings of loneliness. *J. Gerontol. Ser. B* **71** 1049–1058.
- THOMAS, K. S. and MOR, V. (2013). Providing more home-delivered meals is one way to keep older adults with low care needs out of nursing homes. *Health Aff.* **32** 1796–1802.
- THOMAS, K. S., GADBOIS, E. A., SHIELD, R. R., AKOBUNDU, U., MORRIS, A. M. and DOSA, D. M. (2018a). “It’s not just a simple meal. It’s so much more”: Interactions between meals on wheels clients and drivers. *J. Appl. Gerontol.* **39** 151–158.
- THOMAS, K. S., PARIKH, R. B., ZULLO, A. R. and DOSA, D. (2018b). Home-delivered meals and risk of self-reported falls: Results from a randomized trial. *J. Appl. Gerontol.* **37** 41–57. <https://doi.org/10.1177/0733464816675421>
- WINKLER, W. E. (1988). Using the EM algorithm for weight computation in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* 667–671.
- WINKLER, W. (1989). Near automatic weight computation in the Fellegi–Sunter model of record linkage.
- WINKLER, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* 354–359.
- WINKLER, W. E. (1993). Improved decision rules in the Fellegi–Sunter model of record linkage.
- WORTMAN, J. H. and REITER, J. P. (2018). Simultaneous record linkage and causal inference with propensity score subclassification. *Stat. Med.* **37** 3533–3546. [MR3862902 https://doi.org/10.1002/sim.7911](https://doi.org/10.1002/sim.7911)
- ZANELLA, G. (2020). Informed proposals for local MCMC in discrete spaces. *J. Amer. Statist. Assoc.* **115** 852–865. [MR4107684 https://doi.org/10.1080/01621459.2019.1585255](https://doi.org/10.1080/01621459.2019.1585255)

ACCOUNTING FOR SMOKING IN FORECASTING MORTALITY AND LIFE EXPECTANCY

BY YICHENG LI^{*} AND ADRIAN E. RAFTERY[†]

Department of Statistics, University of Washington, ^{}yl83@uw.edu; [†]raftery@uw.edu*

Smoking is one of the main risk factors that has affected human mortality and life expectancy over the past century. Smoking accounts for a large part of the nonlinearities in the growth of life expectancy and of the geographic and gender differences in mortality. As Bongaarts (*Popul. Dev. Rev.* **32** (2006) 605–628) and Janssen (*Genus* **74** (2018) 21) suggested, accounting for smoking could improve the quality of mortality forecasts due to the predictable nature of the smoking epidemic. We propose a new Bayesian hierarchical model to forecast life expectancy at birth for both genders and for 69 countries/regions with good data on smoking-related mortality. The main idea is to convert the forecast of the nonsmoking life expectancy at birth (i.e., life expectancy at birth removing the smoking effect) into life expectancy forecast through the use of the age-specific smoking attributable fraction (ASSAF). We introduce a new age-cohort model for the ASSAF and a Bayesian hierarchical model for nonsmoking life expectancy at birth. The forecast performance of the proposed method is evaluated by out-of-sample validation compared with four other commonly used methods for life expectancy forecasting. Improvements in forecast accuracy and model calibration based on the new method are observed.

REFERENCES

- ALKEMA, L., RAFTERY, A. E., GERLAND, P., CLARK, S. J., PELLETIER, F., BUETTNER, T. and HEILIG, G. K. (2011). Probabilistic projections of the total fertility rate for all countries. *Demography* **48** 815–839.
- BONGAARTS, J. (2006). How long will we live? *Popul. Dev. Rev.* **32** 605–628.
- BONGAARTS, J. (2014). Trends in causes of death in low-mortality countries: Implications for mortality projections. *Popul. Dev. Rev.* **40** 189–212.
- BOOTH, H., HYNDMAN, R. J., TICKLE, L. and DE JONG, P. (2006). Lee–Carter mortality forecasting: A multi-country comparison of variants and extensions. *Demogr. Res.* **15** 289–310.
- BRITTON, J. (2017). Death, disease, and tobacco. *Lancet* **389** 1861–1862.
- BURNS, D. M., LEE, L., SHEN, L. Z., GILPIN, E., TOLLEY, H. D., VAUGHN, J., SHANKS, T. G. et al. (1997). Cigarette smoking behavior in the United States. In *Changes in Cigarette-Related Disease Risks and Their Implication for Prevention and Control. Smoking and Tobacco Control Monograph* **8** 13–42.
- FOKAS, N. (2007). Growth functions, social diffusion, and social change. *Rev. Sociol.* **13** 5–30.
- 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](https://doi.org/10.1198/016214506000001437)
- GODWIN, J. and RAFTERY, A. E. (2017). Bayesian projection of life expectancy accounting for the HIV/AIDS epidemic. *Demogr. Res.* **37** 1549–1610. <https://doi.org/10.4054/DemRes.2017.37.48>
- GRÜBLER, A., NAKIĆENOVIĆ, N. and VICTOR, D. G. (1999). Dynamics of energy technologies and global change. *Energy Policy* **27** 247–280.
- HYNDMAN, R. J. and SHAHID ULLAH, MD. (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](https://doi.org/10.1016/j.csda.2006.07.028)
- HYNDMAN, R. J., BOOTH, H., TICKLE, L. and MAINDONALD, J. (2019). Demography: Forecasting mortality, fertility, migration and population data. R package version 1.22. <https://CRAN.R-project.org/package=demography>.
- ISLAMI, F., TORRE, L. A. and JEMAL, A. (2015). Global trends of lung cancer mortality and smoking prevalence. *Transl. Lung Cancer Res.* **4** 327–338. <https://doi.org/10.3978/j.issn.2218-6751.2015.08.04>

Key words and phrases. Smoking attributable fraction, Bayesian hierarchical model, life expectancy, probabilistic projection.

- JANSSEN, F. (2018). Advances in mortality forecasting: Introduction. *Genus* **74** 21. <https://doi.org/10.1186/s41118-018-0045-7>
- JANSSEN, F. and VAN POPPEL, F. (2015). The adoption of smoking and its effect on the mortality gender gap in Netherlands: A historical perspective. *Biomed. Res. Int.* **2015** 370274. <https://doi.org/10.1155/2015/370274>
- JANSSEN, F., VAN WISSEN, L. J. and KUNST, A. E. (2013). Including the smoking epidemic in internationally coherent mortality projections. *Demography* **50** 1341–1362.
- KING, G. and SONEJI, S. (2011). The future of death in America. *Demogr. Res.* **25** 1–38. <https://doi.org/10.4054/DemRes.2011.25.1>
- KONG, K. A., JUNG-CHOI, K.-H., LIM, D., LEE, H. A., LEE, W. K., BAIK, S. J., PARK, S. H. and PARK, H. (2016). Comparison of prevalence-and smoking impact ratio-based methods of estimating smoking-attributable fractions of deaths. *J. Epidemiol.* **26** 145–154.
- KUCHARAVY, D. and DE GUIO, R. (2011). Logistic substitution model and technological forecasting. *Proc. Eng.* **9** 402–416.
- LEE, R. D. and CARTER, L. R. (1992). Modeling and forecasting US mortality. *J. Amer. Statist. Assoc.* **87** 659–671.
- LEE, R. D. and MILLER, T. (2001). Evaluating the performance of the Lee–Carter method for forecasting mortality. *Demography* **38** 537–549.
- LI, N. and LEE, R. D. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee–Carter method. *Demography* **42** 575–594.
- LI, Y. and RAFTERY, A. E. (2020). Estimating and forecasting the smoking-attributable mortality fraction for both genders jointly in over 60 countries. *Ann. Appl. Stat.* **14** 381–408. [MR4085098 https://doi.org/10.1214/19-AOAS1306](https://doi.org/10.1214/19-AOAS1306)
- LI, Y. and RAFTERY, A. E. (2021). Supplement to “Accounting for smoking in forecasting mortality and life expectancy.” <https://doi.org/10.1214/20-AOAS1381SUPP>
- MARCHETTI, C., MEYER, P. S. and AUSUBEL, J. H. (1996). Human population dynamics revisited with the logistic model: How much can be modeled and predicted? *Technol. Forecast. Soc. Change* **52** 1–30. [https://doi.org/10.1016/0040-1625\(96\)00001-7](https://doi.org/10.1016/0040-1625(96)00001-7)
- MEYER, P. (1994). Bi-logistic growth. *Technol. Forecast. Soc. Change* **47** 89–102.
- OEPPE, J. and VAUPEL, J. W. (2002). Broken limits to life expectancy. *Science* **296** 1029–1031.
- PARASCANDOLA, M. and XIAO, L. (2019). Tobacco and the lung cancer epidemic in China. *Transl. Lung Cancer Res.* **8** S21–S30. <https://doi.org/10.21037/tlcr.2019.03.12>
- PEDROZA, C. (2006). A Bayesian forecasting model: Predicting US male mortality. *Biostatistics* **7** 530–550.
- PETO, R., LOPEZ, A. D., BOREHAM, J., THUN, M. and HEATH, C. (1992). Mortality from tobacco in developed countries: Indirect estimation from national vital statistics. *Lancet* **339** 1268–1278.
- PRESTON, S., HEUVELINE, P. and GUILLOT, M. (2000). *Demography: Measuring and Modeling Population Processes*. Wiley-Blackwell.
- PRESTON, S. H. and WANG, H. (2006). Sex mortality differences in the United States: The role of cohort smoking patterns. *Demography* **43** 631–646.
- RAFTERY, A. E., ALKEMA, L. and GERLAND, P. (2014). Bayesian population projections for the United Nations. *Statist. Sci.* **29** 58–68. [MR3201847 https://doi.org/10.1214/13-STS419](https://doi.org/10.1214/13-STS419)
- RAFTERY, A. E., LALIC, N. and GERLAND, P. (2014). Joint probabilistic projection of female and male life expectancy. *Demogr. Res.* **30** 795–822. <https://doi.org/10.4054/DemRes.2014.30.27>
- RAFTERY, A. E., LI, N., ŠEVČÍKOVÁ, H., GERLAND, P. and HEILIG, G. K. (2012). Bayesian probabilistic population projections for all countries. *Proc. Natl. Acad. Sci. USA* **109** 13915–13921.
- RAFTERY, A. E., CHUNN, J. L., GERLAND, P. and ŠEVČÍKOVÁ, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography* **50** 777–801. <https://doi.org/10.1007/s13524-012-0193-x>
- REITSMA, M. B., FULLMAN, N., NG, M., SALAMA, J. S., ABAJOBIR, A., ABATE, K. H., ABBAFATI, C., ABERA, S. F., ABRAHAM, B. et al. (2017). Smoking prevalence and attributable disease burden in 195 countries and territories, 1990–2015: A systematic analysis from the global burden of disease study 2015. *Lancet* **389** 1885–1906.
- RENSHAW, A. E. and HABERMAN, S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance Math. Econom.* **38** 556–570.
- ROSTRON, B. L. and WILMOTH, J. R. (2011). Estimating the effect of smoking on slowdowns in mortality declines in developed countries. *Demography* **48** 461–479.
- ŠEVČÍKOVÁ, H., LI, N. and GERLAND, P. (2019). MortCast: Estimation and projection of age-specific mortality rates. R package version 2.1-1. <https://CRAN.R-project.org/package=MortCast>.
- ŠEVČÍKOVÁ, H., RAFTERY, A. and CHUNN, J. (2019). bayesLife: Bayesian projection of life expectancy. R package version 4.0-2. <https://CRAN.R-project.org/package=bayesLife>.

- ŠEVČÍKOVÁ, H., ŠEVČÍKOVÁ, H., LI, N., KANTOROVÁ, V., GERLAND, P. and RAFTERY, A. E. (2016). Age-specific mortality and fertility rates for probabilistic population projections. In *Dynamic Demographic Analysis. Springer Ser. Demogr. Methods Popul. Anal.* **39** 285–310. Springer, Cham. MR3525591 https://doi.org/10.1007/978-3-319-26603-9_15
- 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>
- STAN DEVELOPMENT TEAM (2018). RStan: The R interface to Stan. R package version 2.18.2. <http://mc-stan.org/>.
- STOELDRAIJER, L., BONNEUX, L., VAN DUIN, C., VAN WISSEN, L. and JANSSEN, F. (2015). The future of smoking-attributable mortality: The case of England & Wales, Denmark and the Netherlands. *Addiction* **110** 336–345.
- TRIAS LLIMÓS, S. and JANSSEN, F. (2019). Gender gaps in life expectancy and alcohol consumption in Eastern Europe. *N-IUSSP*.
- UNITED NATIONS (2015). World population prospects 2015. United Nations.
- UNITED NATIONS (2017a). World population prospects 2017. United Nations. Accessed: Oct. 15, 2018 <http://population.un.org/wpp/Download/Standard/Population/>.
- UNITED NATIONS (2017b). World population prospects: The 2017 revision, methodology of the United Nations population estimates and projections. https://population.un.org/wpp/Publications/Files/WPP2017_Methodology.pdf.
- VIDRA, N., TRIAS-LLIMÓS, S. and JANSSE, F. (2017). Impact of obesity on trends in life expectancy among different European countries, 1975–2012. *Eur. J. Public Health* **27**.
- WANG, H. and PRESTON, S. H. (2009). Forecasting United States mortality using cohort smoking histories. *Proc. Natl. Acad. Sci. USA* **106** 393–398.
- WHELPTON, P. K. (1936). An empirical method of calculating future population. *J. Amer. Statist. Assoc.* **31** 457–473.
- 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.
- WORLD HEALTH ORGANIZATION (2017). Mortality database. Last accessed: Oct. 15, 2018 http://www.who.int/healthinfo/statistics/mortality_rawdata/en/.

HIGH-FIDELITY HURRICANE SURGE FORECASTING USING EMULATION AND SEQUENTIAL EXPERIMENTS

BY MATTHEW PLUMLEE¹, TAYLOR G. ASHER², WON CHANG³ AND MATTHEW V. BILSKIE⁴

¹Industrial Engineering and Management Sciences, Northwestern University, mplumlee@northwestern.edu

²Department of Marine Sciences, University of North Carolina, tgasher@live.unc.edu

³Division of Statistics and Data Science, University of Cincinnati, won.chang@uc.edu

⁴School of Environmental, Civil, Agricultural, and Mechanical Engineering, University of Georgia, mbilskie@uga.edu

Probabilistic hurricane storm surge forecasting using a high-fidelity model has been considered impractical due to the overwhelming computational expense to run thousands of simulations. This article demonstrates that modern statistical tools enable good forecasting performance using a small number of carefully chosen simulations. This article offers algorithms that quickly handle the massive output of a surge model while addressing the missing data at unsubmerged locations. Also included is a new optimal design criterion for selecting simulations that accounts for the log transform required to statistically model surge data. Hurricane Michael (2018) is used as a testbed for this investigation and provides evidence for the approach's efficacy in comparison to the existing probabilistic surge forecast method.

REFERENCES

- ALTAFF, M., BUTLER, T., MAYO, T., LUO, X., DAWSON, C., HEEMINK, A. and HOTEIT, I. (2014). A comparison of ensemble Kalman filters for storm surge assimilation. *Mon. Weather Rev.* **142** 2899–2914.
- ASHER, T. G., LUETTICH, R. A. JR., FLEMING, J. G. and BLANTON, B. O. (2019). Low frequency water level correction in storm surge models using data assimilation. *Ocean Model.* **144** 101483.
- BECK, J. and GUILLAS, S. (2016). Sequential design with mutual information for computer experiments (MICE): Emulation of a tsunami model. *SIAM/ASA J. Uncertain. Quantificat.* **4** 739–766. [MR3507556](https://doi.org/10.1137/140989613)
- BECT, J., GINSBOURGER, D., LI, L., PICHENY, V. and VAZQUEZ, E. (2012). Sequential design of computer experiments for the estimation of a probability of failure. *Stat. Comput.* **22** 773–793. [MR2909621](https://doi.org/10.1007/s11222-011-9241-4)
- BEVEN, J. L., BERG, R. and HAGEN, A. (2019). Hurricane Michael Tropical Cyclone Report No. AL142018 NOAA/National Hurricane Center.
- BILSKIE, M. V., HAGEN, S. C. and MEDEIROS, S. C. (2019). Unstructured finite element mesh decimation for real-time hurricane storm surge forecasting. *Coastal Eng.* **156** 103622.
- BILSKIE, M. V., COGGIN, D., HAGEN, S. C. and MEDEIROS, S. C. (2015). Terrain-driven unstructured mesh development through semi-automatic vertical feature extraction. *Adv. Water Resour.* **86** 102–118.
- BILSKIE, M. V., HAGEN, S. C., MEDEIROS, S. C., COX, A. T., SALISBURY, M. and COGGIN, D. (2016). Data and numerical analysis of astronomic tides, wind-waves, and hurricane storm surge along the northern Gulf of Mexico. *J. Geophys. Res.* **121** 3625–3658.
- BODE, L. and HARDY, T. A. (1997). Progress and recent developments in storm surge modeling. *J. Hydraul. Eng.* **123** 315–331.
- CANGIALOSI, J. P. (2019). National Hurricane Center Forecast Verification Report: 2018 Hurricane Season Technical Report NOAA/National Hurricane Center.
- CHANG, W., HARAN, M., OLSON, R. and KELLER, K. (2014). Fast dimension-reduced climate model calibration and the effect of data aggregation. *Ann. Appl. Stat.* **8** 649–673. [MR3262529](https://doi.org/10.1214/14-AOAS733)
- CHANG, W., HARAN, M., APPLGATE, P. and POLLARD, D. (2016). Calibrating an ice sheet model using high-dimensional binary spatial data. *J. Amer. Statist. Assoc.* **111** 57–72. [MR3494638](https://doi.org/10.1080/01621459.2015.1108199)

- CHANG, W., KONOMI, B. A., KARAGIANNIS, G., GUAN, Y. and HARAN, M. (2019). Ice model calibration using semi-continuous spatial data. Preprint. Available at [arXiv:1907.13554](https://arxiv.org/abs/1907.13554).
- CONTI, S., GOSLING, J. P., OAKLEY, J. E. and O'HAGAN, A. (2009). Gaussian process emulation of dynamic computer codes. *Biometrika* **96** 663–676. [MR2538764 https://doi.org/10.1093/biomet/asp028](https://doi.org/10.1093/biomet/asp028)
- CRESSIE, N. (1990). The origins of kriging. *Math. Geol.* **22** 239–252. [MR1047810 https://doi.org/10.1007/BF00889887](https://doi.org/10.1007/BF00889887)
- CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, New York. [MR1239641 https://doi.org/10.1002/9781119115151](https://doi.org/10.1002/9781119115151)
- CRESSIE, N., SHI, T. and KANG, E. L. (2010). Fixed rank filtering for spatio-temporal data. *J. Comput. Graph. Statist.* **19** 724–745. [MR2732500 https://doi.org/10.1198/jcgs.2010.09051](https://doi.org/10.1198/jcgs.2010.09051)
- CURRIN, C., MITCHELL, T., MORRIS, M. and YLVISAKER, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Amer. Statist. Assoc.* **86** 953–963. [MR1146343](https://doi.org/10.1198/jcgs.2010.09051)
- CYRIAC, R., DIETRICH, J. C., FLEMING, J. G., BLANTON, B. O., KAISER, C., DAWSON, C. N. and LUETTICH, R. A. (2018). Variability in Coastal Flooding predictions due to forecast errors during Hurricane Arthur. *Coastal Eng.* **137** 59–78.
- 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)
- DAWSON, R. and HALL, J. (2006). Adaptive importance sampling for risk analysis of complex infrastructure systems. *Proc. R. Soc. Ser. A Math. Phys. Eng. Sci.* **462** 3343–3362.
- DIETRICH, J. C., ZIJLEMA, M., WESTERINK, J. J., HOLTHUIJSEN, L. H., DAWSON, C., LUETTICH, R. A. JR., JENSEN, R. E., SMITH, J. M., STELLING, G. S. et al. (2011). Modeling hurricane waves and storm surge using integrally-coupled, scalable computations. *Coastal Eng.* **58** 45–65.
- 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](https://doi.org/10.1198/016214506000001437)
- GRAMACY, R. B. (2016). laGP: Large-scale spatial modeling via local approximate Gaussian processes in R. *J. Stat. Softw.* **72** 1–46.
- GRAMACY, R. B. and LEE, H. K. H. (2009). Adaptive design and analysis of supercomputer experiments. *Technometrics* **51** 130–145. [MR2668170 https://doi.org/10.1198/TECH.2009.0015](https://doi.org/10.1198/TECH.2009.0015)
- GRAMACY, R. B. and LEE, H. K. H. (2012). Cases for the nugget in modeling computer experiments. *Stat. Comput.* **22** 713–722. [MR2909617 https://doi.org/10.1007/s11222-010-9224-x](https://doi.org/10.1007/s11222-010-9224-x)
- GU, M. and BERGER, J. O. (2016). Parallel partial Gaussian process emulation for computer models with massive output. *Ann. Appl. Stat.* **10** 1317–1347. [MR3553226 https://doi.org/10.1214/16-AOAS934](https://doi.org/10.1214/16-AOAS934)
- GU, M. and WANG, L. (2018). Scaled Gaussian stochastic process for computer model calibration and prediction. *SIAM/ASA J. Uncertain. Quantificat.* **6** 1555–1583. [MR3875809 https://doi.org/10.1137/17M1159890](https://doi.org/10.1137/17M1159890)
- GU, M., XIE, F. and WANG, L. (2018). A theoretical framework of the scaled Gaussian stochastic process in prediction and calibration. Preprint. Available at [arXiv:1807.03829](https://arxiv.org/abs/1807.03829).
- HIGDON, D., GATTIKER, J., WILLIAMS, B. and RIGHTLEY, M. (2008). Computer model calibration using high-dimensional output. *J. Amer. Statist. Assoc.* **103** 570–583. [MR2523994 https://doi.org/10.1198/016214507000000888](https://doi.org/10.1198/016214507000000888)
- HUNG, Y., JOSEPH, V. R. and MELKOTE, S. N. (2015). Analysis of computer experiments with functional response. *Technometrics* **57** 35–44. [MR3318347 https://doi.org/10.1080/00401706.2013.869263](https://doi.org/10.1080/00401706.2013.869263)
- JELESNIANSKI, C. P. (1966). Numerical computations of storm surges without bottom stress. *Mon. Weather Rev.* **94** 379–394.
- JELESNIANSKI, C. P., CHEN, J. and SHAFFER, W. A. (1992). SLOSH: Sea, lake, and overland surges from hurricanes Technical Report No. 48 US Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service.
- JIA, G. and TAFLANIDIS, A. A. (2013). Kriging metamodeling for approximation of high-dimensional wave and surge responses in real-time storm/hurricane risk assessment. *Comput. Methods Appl. Mech. Engrg.* **261/262** 24–38. [MR3069864 https://doi.org/10.1016/j.cma.2013.03.012](https://doi.org/10.1016/j.cma.2013.03.012)
- JIA, G., TAFLANIDIS, A. A., NADAL-CARABALLO, N. C., MELBY, J. A., KENNEDY, A. B. and SMITH, J. M. (2016). Surrogate modeling for peak or time-dependent storm surge prediction over an extended coastal region using an existing database of synthetic storms. *Nat. Hazards* **81** 909–938.
- JOHNSON, L. R., GRAMACY, R. B., COHEN, J., MORDECAI, E., MURDOCK, C., ROHR, J., RYAN, S. J., STEWART-IBARRA, A. M. and WEIKEL, D. (2018). Phenomenological forecasting of disease incidence using heteroskedastic Gaussian processes: A Dengue case study. *Ann. Appl. Stat.* **12** 27–66. [MR3773385 https://doi.org/10.1214/17-AOAS1090](https://doi.org/10.1214/17-AOAS1090)

- LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. MR2853727 <https://doi.org/10.1111/j.1467-9868.2011.00777.x>
- LINDLEY, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Stat.* **27** 986–1005. MR0083936 <https://doi.org/10.1214/aoms/1177728069>
- LIU, F. and WEST, M. (2009). A dynamic modelling strategy for Bayesian computer model emulation. *Bayesian Anal.* **4** 393–411. MR2507369 <https://doi.org/10.1214/09-BA415>
- LOEPPKY, J. L., MOORE, L. M. and WILLIAMS, B. J. (2010). Batch sequential designs for computer experiments. *J. Statist. Plann. Inference* **140** 1452–1464. MR2592224 <https://doi.org/10.1016/j.jspi.2009.12.004>
- 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>
- MOREY, S. L., BAIG, S., BOURASSA, M. A., DUKHOVSKOY, D. S. and O'BRIEN, J. J. (2006). Remote forcing contribution to storm-induced sea level rise during Hurricane Dennis. *Geophys. Res. Lett.* **33** L19603.
- PARKER, K., RUGGIERO, P., SERAFIN, K. A. and HILL, D. F. (2019). Emulation as an approach for rapid estuarine modeling. *Coastal Eng.* **150** 79–93.
- PENG, S.-Q. and XIE, L. (2006). Effect of determining initial conditions by four-dimensional variational data assimilation on storm surge forecasting. *Ocean Model.* **14** 1–18.
- PLUMLEE, M. (2017). Bayesian calibration of inexact computer models. *J. Amer. Statist. Assoc.* **112** 1274–1285. MR3735376 <https://doi.org/10.1080/01621459.2016.1211016>
- PLUMLEE, M., ASHER, T. G., CHANG, W. and BILSKIE, M. V (2021). Supplement A to “High-fidelity hurricane surge forecasting using emulation and sequential experiments.” <https://doi.org/10.1214/20-AOAS1398SUPPA>
- PLUMLEE, M., ASHER, T. G., CHANG, W. and BILSKIE, M. V (2021). Supplement B to “High-fidelity hurricane surge forecasting using emulation and sequential experiments.” <https://doi.org/10.1214/20-AOAS1398SUPPB>
- PLUMLEE, M., ASHER, T. G., CHANG, W. and BILSKIE, M. V (2021). Supplement D to “High-fidelity hurricane surge forecasting using emulation and sequential experiments.” <https://doi.org/10.1214/20-AOAS1398SUPPD>
- PLUMLEE, M., ASHER, T. G., CHANG, W. and BILSKIE, M. V (2021). Supplement C to “High-fidelity hurricane surge forecasting using emulation and sequential experiments.” <https://doi.org/10.1214/20-AOAS1398SUPPC>
- RANJAN, P., BINGHAM, D. and MICHAILIDIS, G. (2008). Sequential experiment design for contour estimation from complex computer codes. *Technometrics* **50** 527–541. MR2655651 <https://doi.org/10.1198/0040170080000000541>
- RESIO, D. T. and WESTERINK, J. J. (2008). Modeling the physics of storm surges. *Phys. Today* **61** 33–38.
- ROHMER, J. and IDIER, D. (2012). A meta-modelling strategy to identify the critical offshore conditions for coastal flooding. *Nat. Hazards Earth Syst. Sci.* **12** 2943–2955.
- SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. MR1041765
- SAKIA, R. (1992). The Box-Cox transformation technique: A review. *J. R. Stat. Soc., Ser. D Stat.* **41** 169–178.
- SANSÓ, B. and FOREST, C. (2009). Statistical calibration of climate system properties. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 485–503. MR2750089 <https://doi.org/10.1111/j.1467-9876.2009.00669.x>
- SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2003). *The Design and Analysis of Computer Experiments. Springer Series in Statistics.* Springer, New York. MR2160708 <https://doi.org/10.1007/978-1-4757-3799-8>
- SHEPARD, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference* 517–524. ACM, New York.
- TAFLANIDIS, A. A., KENNEDY, A. B., WESTERINK, J. J., SMITH, J., CHEUNG, K. F., HOPE, M. and TANAKA, S. (2012). Rapid assessment of wave and surge risk during landfalling hurricanes: Probabilistic approach **139** 171–182.
- TANAKA, S., BUNYA, S., WESTERINK, J. J., DAWSON, C. and LUETTICH, R. A. JR. (2011). Scalability of an unstructured grid continuous Galerkin based hurricane storm surge model. *J. Sci. Comput.* **46** 329–358. MR2765498 <https://doi.org/10.1007/s10915-010-9402-1>
- TAYLOR, A. A. and GLAHN, B. (2008). Probabilistic guidance for hurricane storm surge. In *19th Conference on Probability and Statistics* **74** 8.
- TORO, G., NIEDORODA, A., REED, C. and DIVOKY, D. (2010). Quadrature-based approach for the efficient evaluation of surge hazard. *Ocean Eng.* **37** 114–124.
- 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>
- VERNON, I., GOLDSTEIN, M. and BOWER, R. (2014). Galaxy formation: Bayesian history matching for the observable universe. *Statist. Sci.* **29** 81–90. MR3201849 <https://doi.org/10.1214/12-STS412>
- WALKER, A. M., TITLEY, D. W., MANN, M. E., NAJJAR, R. G. and MILLER, S. K. (2018). A fiscally based scale for tropical cyclone storm surge. *Weather Forecast.* **33** 1709–1723.

- 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.
- WESTERINK, J. J., LUETTICH, R. A., FEYEN, J. C., ATKINSON, J. H., DAWSON, C., ROBERTS, H. J., POWELL, M. D., DUNION, J. P., KUBATKO, E. J. et al. (2008). A basin- to channel-scale unstructured grid hurricane storm surge model applied to southern Louisiana. *Mon. Weather Rev.* **136** 833–864.
- WILLMOTT, C. J. and MATSUURA, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **30** 79–82.

MODEL-BASED FEATURE SELECTION AND CLUSTERING OF RNA-SEQ DATA FOR UNSUPERVISED SUBTYPE DISCOVERY

BY DAVID K. LIM^{*}, NAIM U. RASHID[†] AND JOSEPH G. IBRAHIM[‡]

Department of Biostatistics, University of North Carolina at Chapel Hill, ^{*}deelim@live.unc.edu; [†]naim@unc.edu; [‡]ibrahim@bios.unc.edu

Clustering is a form of unsupervised learning that aims to uncover latent groups within data based on similarity across a set of features. A common application of this in biomedical research is in delineating novel cancer subtypes from patient gene expression data, given a set of informative genes. However, it is typically unknown a priori what genes may be informative in discriminating between clusters and what the optimal number of clusters are. Few methods exist for performing unsupervised clustering of RNA-seq samples, and none currently adjust for between-sample global normalization factors, select cluster-discriminatory genes or account for potential confounding variables during clustering. To address these issues, we propose the feature selection and clustering of RNA-seq (FSCseq): a model-based clustering algorithm that utilizes a finite mixture of regression (FMR) model and the quadratic penalty method with a smoothly clipped absolute deviation (SCAD) penalty. The maximization is done by a penalized Classification EM algorithm, allowing us to include normalization factors and confounders in our modeling framework. Given the fitted model, our framework allows for subtype prediction in new patients via posterior probabilities of cluster membership, even in the presence of batch effects. Based on simulations and real data analysis, we show the advantages of our method relative to competing approaches.

REFERENCES

- AL-KHASAWNEH, M. F. (2010). Estimating the negative binomial dispersion parameter. *Asian J. Math. Stat.* **3** 1–15. [MR2725773 https://doi.org/10.3923/ajms.2010.1.15](https://doi.org/10.3923/ajms.2010.1.15)
- ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* **11** R106. <https://doi.org/10.1186/gb-2010-11-10-r106>
- ARAN, D., SIROTA, M. and BUTTE, A. J. (2015). Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**. <https://doi.org/10.1038/ncomms9971>
- BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput. Statist. Data Anal.* **41** 561–575. [MR1968069 https://doi.org/10.1016/S0167-9473\(02\)00163-9](https://doi.org/10.1016/S0167-9473(02)00163-9)
- BOURGON, R., GENTLEMAN, R. and HUBER, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. USA* **107** 9546–9551. <https://doi.org/10.1073/pnas.0914005107>
- CARTER, S. L., CIBULSKIS, K., HELMAN, E., MCKENNA, A., SHEN, H., ZACK, T., LAIRD, P. W., ONOFRIO, R. C., WINCKLER, W. et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30** 413–421. <https://doi.org/10.1038/nbt.2203>
- CELEUX, G. and GOVAERT, G. (1990). Stochastic algorithms for clustering. In *COMPSTAT 1990 (Dubrovnik, 1990)* 3–8. Physica, Heidelberg. [MR1118609](https://doi.org/10.1007/978-3-7091-1186-0_1)
- CELEUX, G. and GOVAERT, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Comput. Statist. Data Anal.* **14** 315–332. [MR1192205 https://doi.org/10.1016/0167-9473\(92\)90042-E](https://doi.org/10.1016/0167-9473(92)90042-E)
- CHARRAD, M., GHAZZALI, N., BOITEAU, V. and NIKNAFS, A. (2014). NbClust: AnRPackage for determining the relevant number of clusters in a data set. *J. Stat. Softw.* **61**. <https://doi.org/10.18637/jss.v061.i06>
- CHIA, S. K., BRAMWELL, V. H., TU, D., SHEPHERD, L. E., JIANG, S. et al. (2012). A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. *Clin. Cancer Res.* **18** 4465–4472. <https://doi.org/10.1158/1078-0432.ccr-12-0286>

- CHUNG, N., ZHANG, X. D., KREAMER, A., LOCCO, L., KUAN, P. F. et al. (2008). Median absolute deviation to improve hit selection for genome-scale RNAi screens. *J. Biomol. Screen.* **13** 149–158. <https://doi.org/10.1177/1087057107312035>
- COLAPRICO, A., SILVA, T. C., OLSEN, C., GAROFANO, L., CAVA, C., GAROLINI, D. et al. (2015). TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44** e71–e71. <https://doi.org/10.1093/nar/gkv1507>
- DELLAERT, F. (2002). The expectation maximization algorithm. Technical report, Georgia Institute of Technology.
- 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>
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.
- FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. MR2415737 <https://doi.org/10.1214/07-AOAS131>
- GARCIA, R. I., IBRAHIM, J. G. and ZHU, H. (2010). Variable selection in the Cox regression model with covariates missing at random. *Biometrics* **66** 97–104. MR2756695 <https://doi.org/10.1111/j.1541-0420.2009.01274.x>
- GAUJOUX, R. and SEOIGHE, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinform.* **11** 367. <https://doi.org/10.1186/1471-2105-11-367>
- GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLIER, H., LOH, M. L., DOWNING, J. R. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286** 531–537. <https://doi.org/10.1126/science.286.5439.531>
- GROSSMAN, R. L., HEATH, A. P., FERRETTI, V., VARMUS, H. E., LOWY, D. R., KIBBE, W. A. and STAUDT, L. M. (2016). Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375** 1109–1112. <https://doi.org/10.1056/nejmp1607591>
- GRÜN, D., KESTER, L. and VAN OUDENAARDEN, A. (2014). Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11** 637–640. <https://doi.org/10.1038/nmeth.2930>
- HILBE, J. M. (2009). *Modeling Count Data*. Cambridge Univ. Press, Cambridge. <https://doi.org/10.1017/cbo9781139236065>
- HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67. <https://doi.org/10.1080/00401706.1970.10488634>
- HUSZNO, J. and KOLOSZA, Z. (2019). Molecular characteristics of breast cancer according to clinicopathological factors. *Mol. Clin. Oncol.* **11** 192–200. <https://doi.org/10.3892/mco.2019.1869>
- INOUE, D. I., YANG, E., ALLEN, G. I. and RAVIKUMAR, P. (2017). A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdiscip. Rev.: Comput. Stat.* **9** e1398, 25. MR3648601 <https://doi.org/10.1002/wics.1398>
- JASKOWIAK, P. A., COSTA, I. G. and CAMPELLO, R. J. G. B. (2018). Clustering of RNA-seq samples: Comparison study on cancer data. *Methods* **132** 42–49. <https://doi.org/10.1016/j.ymeth.2017.07.023>
- JOHNSON, W. E., LI, C. and RABINOVIC, A. (2006). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8** 118–127. <https://doi.org/10.1093/biostatistics/kxj037>
- KLEIN, R. W. and DUBES, R. C. (1989). Experiments in projection and clustering by simulated annealing. *Pattern Recognit.* **22** 213–220. [https://doi.org/10.1016/0031-3203\(89\)90067-8](https://doi.org/10.1016/0031-3203(89)90067-8)
- KOBOLDT, D. C., FULTON, R. S., MCLELLAN, M. D., SCHMIDT, H., KALICKIVEIZER, J. et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70. <https://doi.org/10.1038/nature11412>
- LEEK, J. T. (2014). svaseq: Removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42** e161–e161. <https://doi.org/10.1093/nar/gku864>
- LEEK, J. T., SCHARPF, R. B., BRAVO, H. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GEMAN, D., BAGGERLY, K. and IRIZARRY, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11** 733–739. <https://doi.org/10.1038/nrg2825>
- LI, B. and DEWEY, C. N. (2011). RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinform.* **12** 323. <https://doi.org/10.1186/1471-2105-12-323>
- LI, P., PIAO, Y., SHON, H. S. and RYU, K. H. (2015). Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-seq data. *BMC Bioinform.* **16** 347. <https://doi.org/10.1186/s12859-015-0778-7>
- LI, Q., NOEL-MACDONNELL, J. R., KOESTLER, D. C., GOODE, E. L. and FRIDLEY, B. L. (2018). Subject level clustering using a negative binomial model for small transcriptomic studies. *BMC Bioinform.* **19** 474. <https://doi.org/10.1186/s12859-018-2556-9>

- LIM, D. K., RASHID, N. U. and IBRAHIM, J. G. (2021a). Supplement A: Supporting information for model-based feature selection and clustering of RNA-seq data for unsupervised subtype discovery. <https://doi.org/10.1214/20-AOAS1407SUPPA>
- LIM, D. K., RASHID, N. U. and IBRAHIM, J. G. (2021b). Supplement B: FSCseq R package. <https://doi.org/10.1214/20-AOAS1407SUPPB>
- LIM, D. K., RASHID, N. U. and IBRAHIM, J. G. (2021c). Supplement C: FSCseq reproducibility code. <https://doi.org/10.1214/20-AOAS1407SUPPC>
- LOVE, M. I., HUBER, W. and ANDERS, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15** 550. <https://doi.org/10.1186/s13059-014-0550-8>
- MAO, J.-H., VAN DIEST, P. J., PEREZ-LOSADA, J. and SNIJDERS, A. M. (2017). Revisiting the impact of age and molecular subtype on overall survival after radiotherapy in breast cancer patients. *Sci. Rep.* **7** 12587. <https://doi.org/10.1038/s41598-017-12949-5>
- MCLACHLAN, G. J. and KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR2392878 <https://doi.org/10.1002/9780470191613>
- MENG, X.-L. (1994). On the rate of convergence of the ECM algorithm. *Ann. Statist.* **22** 326–339. MR1272086 <https://doi.org/10.1214/aos/1176325371>
- MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503 <https://doi.org/10.1093/biomet/80.2.267>
- MO, Q. and SHEN, R. (2019). iClusterPlus: Integrative clustering of multi-type genomic data. R package version 1.20.0.
- MO, Q., WANG, S., SESHAN, V. E., OLSHEN, A. B., SCHULTZ, N. et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. USA* **110** 4245–4250. <https://doi.org/10.1073/pnas.1208949110>
- MOOHTA, V. K., LINDGREN, C. M., ERIKSSON, K.-F., SUBRAMANIAN, A., SIHAG, S., LEHAR, J., PUIGSERVER, P., CARLSSON, E., RIDDERSTRÅLE, M. et al. (2003). PGC-1-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34** 267–273. <https://doi.org/10.1038/ng1180>
- NEAL, R. and HINTON, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models* 355–368. Kluwer Academic, Dordrecht.
- NOCEDAL, J. and WRIGHT, S. J. (2000). *Numerical Optimization. Springer Series in Operations Research*. Springer, New York. MR1713114 <https://doi.org/10.1007/b98874>
- NOEL-MACDONNELL, J. R., USSET, J., GOODE, E. L. and FRIDLEY, B. L. (2018). Assessment of data transformations for model-based clustering of RNA-seq data. *PLoS ONE* **13** e0191758. <https://doi.org/10.1371/journal.pone.0191758>
- PAN, W. and SHEN, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* **8** 1145–1164.
- PAN, W., SHEN, X. and LIU, B. (2013). Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *J. Mach. Learn. Res.* **14** 1865–1889. MR3104498
- PATRO, R., DUGGAL, G., LOVE, M. I., IRIZARRY, R. A. and KINGSFORD, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14** 417–419. <https://doi.org/10.1038/nmeth.4197>
- PEROU, C. M., SØRLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S. et al. (2000). Molecular portraits of human breast tumours. *Nature* **406** 747–752. <https://doi.org/10.1038/35021093>
- PICORNELL, A. C., ECHAVARRIA, I., ALVAREZ, E., LÓPEZ-TARRUELLA, S., JEREZ, Y., HOADLEY, K., PARKER, J. S., DEL MONTE-MILLÁN, M., RAMOS-MEDINA, R. et al. (2019). Breast cancer PAM50 signature: Correlation and concordance between RNA-seq and digital multiplexed gene expression technologies in a triple negative breast cancer series. *BMC Genomics* **20** 452. <https://doi.org/10.1186/s12864-019-5849-0>
- PIEGORSCH, W. W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics* **46** 863–867. MR1085821 <https://doi.org/10.2307/2532104>
- RAJ-KUMAR, P.-K., LIU, J., HOOKE, J. A., KOVATICH, A. J., KVECER, L., SHRIVER, C. D. and HU, H. (2019). PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. *Sci. Rep.* **9** 7956. <https://doi.org/10.1038/s41598-019-44339-4>
- REESE, S. E., ARCHER, K. J., THERNEAU, T. M., ATKINSON, E. J., VACHON, C. M., DE ANDRADE, M., KOCHER, J.-P. A. and ECKEL-PASSOW, J. E. (2013). A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics* **29** 2877–2883. <https://doi.org/10.1093/bioinformatics/btt480>
- REICHART, R. and RAPPOPORT, A. (2009). The NVI clustering evaluation measure. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning. CoNLL '09* 165–173. Association for Computational Linguistics, Stroudsburg, PA, USA.

- REYNOLDS, A. P., RICHARDS, G. and RAYWARD-SMITH, V. J. (2004). The application of K-medoids and PAM to the clustering of rules. In *Intelligent Data Engineering and Automated Learning—IDEAL 2004, 5th International Conference* 173–178.
- 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. <https://doi.org/10.1038/s41467-017-02554-5>
- ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, G. K. (2009). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- ROSE, K. (1998). Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proc. IEEE* **86** 2210–2239. <https://doi.org/10.1109/5.726788>
- SCHURCH, N. J., SCHOFIELD, P., GIERLIŃSKI, M., COLE, C., SHERSTNEV, A., SINGH, V., WROBEL, N., GHARBI, K., SIMPSON, G. G. et al. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22** 839–851. <https://doi.org/10.1261/rna.053959.115>
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. [MR0468014](https://doi.org/10.1214/aos/1176344946)
- SCRUCCA, L., FOP, M., MURPHY, T. B. and RAFTERY, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8** 289–317.
- SI, Y., LIU, P., LI, P. and BRUTNELL, T. P. (2013). Model-based clustering for RNA-seq data. *Bioinformatics* **30** 197–205. <https://doi.org/10.1093/bioinformatics/btt632>
- SMID, M., WANG, Y., ZHANG, Y., SIEUWERTS, A. M., YU, J., KLIJN, J. G. M., FOEKENS, J. A. and MARTENS, J. W. M. (2008). Subtypes of breast cancer show preferential site of relapse. *Cancer Res.* **68** 3108–3114. <https://doi.org/10.1158/0008-5472.can-07-5644>
- STREHL, A. and GHOSH, J. (2003). Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3** 583–617. [MR1991087 https://doi.org/10.1162/153244303321897735](https://doi.org/10.1162/153244303321897735)
- SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102** 15545–15550. <https://doi.org/10.1073/pnas.0506580102>
- TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 411–423. [MR1841503 https://doi.org/10.1111/1467-9868.00293](https://doi.org/10.1111/1467-9868.00293)
- TOWNES, F. W., HICKS, S. C., ARYEE, M. J. and IRIZARRY, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model. *Genome Biol.* **20** 295. <https://doi.org/10.1186/s13059-019-1861-6>
- TRITCHLER, D., PARKHOMENKO, E. and BEYENE, J. (2009). Filtering genes for cluster and network analysis. *BMC Bioinform.* **10** 193. <https://doi.org/10.1186/1471-2105-10-193>
- TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* **109** 475–494. [MR1835069 https://doi.org/10.1023/A:1017501703105](https://doi.org/10.1023/A:1017501703105)
- VAN LAARHOVEN, P. J. M. and AARTS, E. H. L. (1987). *Simulated Annealing: Theory and Applications. Mathematics and Its Applications* **37**. D. Reidel Publishing Co., Dordrecht. [MR0904050 https://doi.org/10.1007/978-94-015-7744-1](https://doi.org/10.1007/978-94-015-7744-1)
- VIDMAN, L., KÄLLBERG, D. and RYDÉN, P. (2019). Cluster analysis on high dimensional RNA-seq data with applications to cancer research—An evaluation study. *PLoS ONE* **14** e0219102. <https://doi.org/10.1371/journal.pone.0219102>
- VINH, N. X., EPPS, J. and BAILEY, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11** 2837–2854. [MR2738784](https://doi.org/10.1162/109974610X505543)
- WU, T. T. and LANGE, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* **2** 224–244. [MR2415601 https://doi.org/10.1214/07-AOAS147](https://doi.org/10.1214/07-AOAS147)
- WU, A. R., NEFF, N. F., KALISKY, T., DALERBA, P., TREUTLEIN, B., ROTHENBERG, M. E., MBURU, F. M., MANTALAS, G. L., SIM, S. et al. (2013). Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11** 41–46. <https://doi.org/10.1038/nmeth.2694>
- YANG, F., DING, P. and HUANG, R. (2015). Clinicopathological significance and potential drug target of CDH1 in breast cancer: A meta-analysis and literature review. *Drug Des. Dev. Ther.* **9** 5277. <https://doi.org/10.2147/dddt.s86929>
- YANG, K., GAO, J. and LUO, M. (2019). Identification of key pathways and hub genes in basal-like breast cancer using bioinformatics analysis. *OncoTargets Ther.* **12** 1319–1331. <https://doi.org/10.2147/ott.s158619>
- YANG, L., SHEN, Y., YUAN, X., ZHANG, J. and WEI, J. (2017). Analysis of breast cancer subtypes by AP-ISA biclustering. *BMC Bioinform.* **18** 481. <https://doi.org/10.1186/s12859-017-1926-z>

- YERSAL, O. and BARUTCA, S. (2014). Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J. Clin. Oncol.* **5** 412–424. <https://doi.org/10.5306/wjco.v5.i3.412>
- ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. [MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x)
- ZWIENER, I., FRISCH, B. and BINDER, H. (2014). Transforming RNA-seq data to improve the performance of prognostic gene signatures. *PLoS ONE* **9** e85150. <https://doi.org/10.1371/journal.pone.0085150>

AN EMPIRICAL BAYES CHANGE-POINT MODEL FOR TRANSCRIPTOME TIME-COURSE DATA

BY TIAN TIAN^{1,*}, RUIHUA CHENG² AND ZHI WEI^{1,†}

¹Department of Computer Science, New Jersey Institute of Technology, *tt72@njit.edu; †zhiwei@njit.edu

²Big Data Statistics Research Center, Tianjin University of Finance and Economics, r.cheng665@gmail.com

Time-course experiments are commonly conducted to capture temporal changes. It is generally of interest to detect if any changes happen over time, which we define as a detection problem. If there is a change, it is informative to know when the change is, which we define as an identification problem. It is often desired to control Type I error rate at a nominal level while applying a testing procedure to detect or identify these changes. Quite a few analytic methods have been proposed. Most existing methods aim to solve either the detection problem or, more recently, the identification problem. Here, we propose to solve these two problems using a unified multiple-testing framework built upon an empirical Bayes change-point model. Our model provides a flexible framework that can account for sophisticated temporal gene expression patterns. We show that our testing procedure is valid and asymptotically optimal in the sense of rejecting the maximum number of null hypotheses, while the Bayesian false discovery rate (FDR) can be controlled at a predefined nominal level. Simulation studies and application to real transcriptome time-course data illustrate that our proposed model is a flexible and powerful method to capture various temporal patterns in analysis of time-course data.

REFERENCES

- ABADI, M., BARHAM, P., CHEN, J., CHEN, Z., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., IRVING, G. et al. (2016). TensorFlow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* 265–283. USENIX Association, Savannah, GA, USA.
- ALLAIRE, J. and TANG, Y. (2020). tensorflow: R Interface to ‘TensorFlow’. R package version 2.0.0.
- ARBEITMAN, M. N., FURLONG, E. E., IMAM, F., JOHNSON, E., NULL, B. H., BAKER, B. S., KRASNOW, M. A., SCOTT, M. P., DAVIS, R. W. et al. (2002). Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297** 2270–2275.
- BAR-JOSEPH, Z., GITTER, A. and SIMON, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* **13** 552–564. <https://doi.org/10.1038/nrg3244>
- 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)
- BARRY, D. and HARTIGAN, J. A. (1993). A Bayesian analysis for change point problems. *J. Amer. Statist. Assoc.* **88** 309–319. [MR1212493](https://doi.org/10.1080/01621459.1993.10477433)
- 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](https://doi.org/10.2307/2346178)
- CALVANO, S. E., XIAO, W., RICHARDS, D. R., FELCIANO, R. M., BAKER, H. V., CHO, R. J., CHEN, R. O., BROWNSTEIN, B. H., COBB, J. P. et al. (2005). A network-based analysis of systemic inflammation in humans. *Nature* **437** 1032–1037.
- DENISON, D. G. T., HOLMES, C. C., MALLICK, B. K. and SMITH, A. F. M. (2002). *Bayesian Methods for Non-linear Classification and Regression*. Wiley Series in Probability and Statistics. Wiley, Chichester. [MR1962778](https://doi.org/10.1002/9781118131224)
- DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. Oxford Statistical Science Series **25**. Oxford Univ. Press, Oxford. [MR2049007](https://doi.org/10.1093/oxfordjbs.oxfordjbs.25.1.1)
- EFRON, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics (IMS) Monographs **1**. Cambridge Univ. Press, Cambridge. [MR2724758 https://doi.org/10.1017/CBO9780511761362](https://doi.org/10.1017/CBO9780511761362)

- EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. [MR1946571](#) <https://doi.org/10.1198/016214501753382129>
- GAUTIER, L., COPE, L., BOLSTAD, B. M. and IRIZARRY, R. A. (2004). affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20** 307–315.
- IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264.
- KALAITZIS, A. A. and LAWRENCE, N. D. (2011). 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>
- KENDZIORSKI, C., NEWTON, M., LAN, H. and GOULD, M. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med.* **22** 3899–3914.
- KIMELDORF, G. S. and WAHBA, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat.* **41** 495–502. [MR0254999](#) <https://doi.org/10.1214/aoms/1177697089>
- KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. Preprint. Available at [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- LAW, C. W., CHEN, Y., SHI, W. and SMYTH, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15** 1.
- LÖNNSTEDT, I., GRANT, S., BEGLEY, G. and SPEED, T. (2005). Microarray analysis of two interacting treatments: A linear model and trends in expression over time.
- LOVE, M. I., HUBER, W. and ANDERS, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15** 550. <https://doi.org/10.1186/s13059-014-0550-8>
- MURPHY, K. P. (2007). Conjugate Bayesian analysis of the Gaussian distribution. *Def.* **1** 16.
- NEWTON, M. A., KENDZIORSKI, C. M., RICHMOND, C. S., BLATTNER, F. R. and TSUI, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.* **8** 37–52. <https://doi.org/10.1089/106652701300099074>
- ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.
- SMYTH, G. K. (2005). limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* 397–420. Springer, New York, NY.
- SUN, W. and WEI, Z. (2011). Multiple testing for pattern identification, with applications to microarray time-course experiments. *J. Amer. Statist. Assoc.* **106** 73–88. [MR2816703](#) <https://doi.org/10.1198/jasa.2011.ap09587>
- TAI, Y. C. and SPEED, T. P. (2005). *DNA Microarrays. Chapter 20: Statistical Analysis of Microarray Time Course Data*. CRC Press/CRC, New York.
- TAI, Y. C. and SPEED, T. P. (2006). A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Statist.* **34** 2387–2412. [MR2291504](#) <https://doi.org/10.1214/0090536060000000759>
- TIAN, T., CHENG, R. and WEI, Z. (2021a). Supplement to “An empirical bayes change-point model for transcriptome time course data.” <https://doi.org/10.1214/20-AOAS1403SUPPA>
- TIAN, T., CHENG, R. and WEI, Z. (2021b). Source code to “An empirical bayes change-point model for transcriptome time course data.” <https://doi.org/10.1214/20-AOAS1403SUPPB>
- TIAN, B., NOWAK, D. E. and BRASIER, A. R. (2005). A TNF-induced gene expression program under oscillatory NF- κ B control. *BMC Genomics* **6** 1.
- WANG, Z., GERSTEIN, M. and SNYDER, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10** 57–63. <https://doi.org/10.1038/nrg2484>
- WANG, W., WEI, Z. and LI, H. (2014). A change-point model for identifying 3'UTR switching by next-generation RNA sequencing. *Bioinformatics* **30** 2162–2170.
- XIE, J., CAI, T. T., MARIS, J. and LI, H. (2011). Optimal false discovery rate control for dependent data. *Stat. Interface* **4** 417–430. [MR2868825](#) <https://doi.org/10.4310/SII.2011.v4.n4.a1>
- XUAN, X. and MURPHY, K. (2007). Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th International Conference on Machine Learning* 1055–1062. ACM, New York.
- YANG, J., PENFOLD, C. A., GRANT, M. R. and RATTRAY, M. (2016). Inferring the perturbation time from biological time course data. *Bioinformatics* **32** 2956–2964.
- YUAN, M. and KENDZIORSKI, C. (2006). Hidden Markov models for microarray time course data in multiple biological conditions. *J. Amer. Statist. Assoc.* **101** 1323–1332. [MR2307565](#) <https://doi.org/10.1198/016214505000000394>
- ZHANG, J. and WEI, Z. (2016). An empirical Bayes change-point model for identifying 3' and 5' alternative splicing by next-generation RNA sequencing. *Bioinformatics* **32** 1823–1831.

- ZHAO, L. P., PRENTICE, R. and BREEDEN, L. (2001). Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl. Acad. Sci. USA* **98** 5631–5636.
- ZHAO, Z., WANG, W. and WEI, Z. (2013). An empirical Bayes testing procedure for detecting variants in analysis of next generation sequencing data. *Ann. Appl. Stat.* **7** 2229–2248. MR3161720 <https://doi.org/10.1214/13-AOAS660>

The Annals of Applied Statistics

Next Issues

- Probabilistic forecasting of the Arctic Sea ice edge with contour modeling
HANNAH DIRECTOR, ADRIAN RAFTERY AND CECILIA BITZ
- A Hermite–Gaussian based exoplanet radial velocity estimation method
PARKER H. HOLZER, JESSI CISEWSKI-KEHE, DEBRA FISCHER AND LILY ZHAO
- A continuous-time semi-Markov model for animal movement in a dynamic environment
DEVIN JOHNSON, NOEL PELLAND AND JEREMY STERLING
- Prediction of the NASH through penalized mixture of logistic regression models
MARIE MORVAN, ÉMILIE DEVIJVER, MADISON GIACOFI AND VALÉRIE MONBET
- Inference of large modified Poisson-type graphical models: Application to RNA-seq data in
childhood atopic asthma studies RONG ZHANG, ZHAO REN,
JUAN C. CELEDÓN AND WEI CHEN
- Inflection points in community-level homeless rates
CHRISTOPHER GLYNN, THOMAS H. BYRNE AND DENNIS P. CULHANE
- A multivariate spatio-temporal change point model of opioid overdose deaths in Ohio
STACI A. HEPLER, LANCE WALLER AND DAVID KLINE
- Two-way sparsity for time-varying networks, with applications in genomics
THOMAS E. BARTLETT
- Additive stacking for disaggregate electricity demand forecasting CHRISTIAN CAPEZZA,
BIAGIO PALUMBO, YANNIG GOUDE, SIMON N. WOOD AND MATTEO FASIOLO
- Estimating high-resolution Red Sea surface temperature hotspots, using a low-rank
semiparametric spatial model ARNAB HAZRA AND RAPHAEL HUSER
- Efficient Bayesian inference of general Gaussian models on large phylogenetic trees
PAUL BASTIDE, LAM SI TUNG HO, GUY BAELE,
PHILIPPE LEMEY AND MARC A. SUCHARD
- Tensor-variate finite mixture modeling for the analysis of university professor remuneration
SHUCHISMITA SARKAR, VOLODYMYR MELNYKOV AND XUWEN ZHU
- A Bayesian semiparametric Jolly–Seber model with individual heterogeneity: An application to
migratory mallards at stopover GUOHUI WU, SCOTT HOLAN,
ALEXIS AVRIL AND JONAS WALDENSTRÖM
- Scalable penalized spatiotemporal land-use regression for ground-level nitrogen dioxide
KYLE P. MESSIER AND MATTHIAS KATZFUSS
- A compositional model to assess expression changes from single-cell RNA-seq data
XIUYU MA, KEEGAN KORTHAUER,
CHRISTINA KENDZIORSKI AND MICHAEL A. NEWTON
- Unsupervised streaming anomaly detection for instrumented infrastructure
HENRIQUE HELFER HOELTGEBAUM, NIALL ADAMS AND DIN-HOUN LAU
- Function-on-function regression for the identification of epigenetic regions exhibiting windows
of susceptibility to environmental exposures ... MICHELE ZEMPLÉNYI, MARK J. MEYER,
ANDRES CARDENAS, MARIE-FRANCE HIVERT, SHERYL L. RIFAS-SHIMAN,
HEIKE GIBSON, ITAI KLOOG, JOEL SCHWARTZ, EMILY OKEN, DAWN L. DEMEO,
DIANE R. GOLD AND BRENT A. COULL
- Rapid design of metamaterials via multi-target Bayesian optimization
YANG YANG, CHUNLIN JI AND KE DENG

Continued

The Annals of Applied Statistics

Next Issues—Continued

- Causal mediation analysis for sparse and irregular longitudinal data SHUXI ZENG,
STACY ROSENBAUM, SUSAN C. ALBERTS, ELIZABETH A. ARCHIE AND FAN LI
- Length-biased semi-competing risks models for cross-sectional data: An application to current
duration of pregnancy attempt data ALEXANDER MCLAIN, SIYUAN GUO,
MARIE THOMA AND JIAJIA ZHANG
- Two-stage circular-circular regression with zero-inflation: Application to medical sciences
JAYANT JHA AND PRAJAMITRA BHUYAN
- A statistical pipeline for identifying physical features that differentiate classes of 3D shapes
BRUCE WANG, TIMOTHY SUDIJONO, HENRY KIRVESLAHTI, TINGRAN GAO,
DOUGLAS M. BOYER, SAYAN MUKHERJEE AND LORIN CRAWFORD
- Large scale multiple inference of collective dependence with applications to protein function
WEN ZHOU, ZHAO REN, ROBERT JERNIGAN AND KEJUE JIA
- A covariance-enhanced approach to multi-tissue joint eQTL mapping with application to
transcriptome-wide association studies AARON J. MOLSTAD, WEI SUN AND LI HSU
- Joint and individual analysis of breast cancer histologic images and genomic covariates
IAIN CARMICHAEL, BENJAMIN C. CALHOUN, KATHERINE A. HOADLEY,
MELISSA A. TROESTER, JOSEPH GERADTS, HEATHER D. COUTURE, LINNEA OLSSON,
CHARLES M. PEROU, MARC NIETHAMMER, JAN HANNIG AND J. S. MARRON
- A Bayesian nonparametric model for inferring subclonal populations from structured DNA
sequencing data SHAI HE, AARON SCHEIN,
VISHAL SARSANI AND PATRICK FLAHERTY
- Perturbed factor analysis: Accounting for group differences in exposure profiles
ARKAPRAVA ROY, ISAAC LAVINE, AMY H. HERRING AND DAVID B. DUNSON
- Global estimation and scenario-based projections of sex ratio at birth and missing female births
using a Bayesian hierarchical time series mixture model FENGQING CHAO,
PATRICK GERLAND, ALEX R. COOK AND LEONTINE ALKEMA
- A bridging model to reconcile statistics based on data from multiple surveys
ANDREEA LUISA ERCIULESCU
- Targeted smooth Bayesian causal forests: An analysis of heterogeneous treatment effects for
simultaneous versus interval medical abortion regimens over gestation
JENNIFER E. STARLING, JARED S. MURRAY, PATRICIA A. LOHR,
ABIGAIL R. A. AIKEN, CARLOS M. CARVALHO AND JAMES G. SCOTT
- Partial-mastery cognitive diagnosis models
ZHUORAN SHANG, ELENA A. EROSHEVA AND GONGJUN XU
- Simultaneous inference of periods and period-luminosity relations for Mira variable stars
SHIYUAN HE, ZHENFENG LIN, WENLONG YUAN,
LUCAS M. MACRI AND JIANHUA Z. HUANG
- A Bayesian nonparametric approach to super-resolution single-molecule localization
MARIANO IGNACIO GABITTO, HERVE MARIE-NELLY, ARI PAKMAN, ANDRAS PATAKI,
XAVIER DARZACQ AND MICHAEL JORDAN
- Zero-inflated quantile rank-score based test (ZIQRank) with application to scRNA-seq
differential gene expression analysis WODAN LING, WENFEI ZHANG,
BIN CHENG AND YING WEI

Continued

The Annals of Applied Statistics

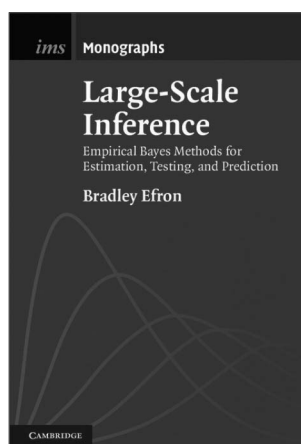
Next Issues—*Continued*

- Hierarchical integrated spatial process modeling of monotone West Antarctic snow density curves PHILIP A. WHITE, DURBAN G. KEELER AND SUMMER RUPPER
- Sparse matrix linear models for structured high-throughput data
JANE W. LIANG AND SAUNAK SEN
- Modeling past event feedback through biomarker dynamics in the multi-state event analysis for cardiovascular disease data CHUOXIN MA, HONGSHENG DAI AND JIANXIN PAN
- Estrogen receptor expression on breast cancer patients' survival under shape restricted Cox regression model JING QIN, GENG DENG, JING NING, AO YUAN AND YU SHEN
- Markov random field models for vector-based representations of landscapes
PATRIZIA ZAMBERLETTI, JULIEN PAPAIX, EDITH GABRIEL AND THOMAS OPITZ
- The information in covariate imbalance in studies of hormone replacement therapy
DYLAN SMALL, RUOQI YU AND PAUL ROSENBAUM



The Institute of Mathematical Statistics presents

IMS MONOGRAPHS



Large-Scale Inference: *Empirical Bayes Methods for* *Estimation, Testing, and Prediction*

Bradley Efron

We live in a new age for statistical inference, where modern scientific technology such as microarrays and fMRI machines routinely produce thousands and sometimes millions of parallel data sets, each with its own estimation or testing problem. Doing thousands of problems at once is more than repeated application of classical methods. Taking an empirical Bayes approach, Bradley Efron, inventor of the bootstrap, shows how information accrues across problems in a way that combines Bayesian and frequentist ideas. Estimation, testing, and prediction blend in this framework, producing opportunities for new methodologies of increased power. New difficulties also arise, easily leading to flawed inferences. This book takes a careful look at both the promise and pitfalls of large-scale statistical inference, with particular attention to false discovery rates, the most successful of the new statistical techniques. Emphasis is on the inferential ideas underlying technical developments, illustrated using a large number of real examples.

MS member? Claim
your 40% discount:
www.cambridge.org/ims

Paperback price
US\$23.99
(non-member price
\$39.99)

www.cambridge.com/ims

Cambridge University Press, in conjunction with the Institute of Mathematical Statistics, established the *IMS Monographs* and *IMS Textbooks* series of high-quality books. The series editors are Xiao-Li Meng, Susan Holmes, Ben Hambly, D. R. Cox and Alan Agresti.