

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

- Hierarchical multidimensional scaling for the comparison of musical performance styles ANNA K. YANCHENKO AND PETER D. HOFF 1581
- Measuring timeliness of annual reports filing by jump additive models . . YICHENG KANG 1604
- Monotonic effects of characteristics on returns
JARED D. FISHER, DAVID W. PUELZ AND CARLOS M. CARVALHO 1622
- Data fusion model for speciated nitrogen to identify environmental drivers and improve estimation of nitrogen in lakes ERIN M. SCHLIEP, SARAH M. COLLINS, SHIRLEY ROJAS-SALAZAR, NOAH R. LOTTIG AND EMILY H. STANLEY 1651
- Feature selection for data integration with mixed multiview data
YULIA BAKER, TIFFANY M. TANG AND GENEVERA I. ALLEN 1676
- Mining events with declassified diplomatic documents YUANJUN GAO, JACK GOETZ, MATTHEW CONNELLY AND RAHUL MAZUMDER 1699
- Classification from only positive and unlabeled functional data
YOSHIKAZU TERADA, ISSEI OGASAWARA AND KEN NAKATA 1724
- Identifying main effects and interactions among exposures using Gaussian processes FEDERICO FERRARI AND DAVID B. DUNSON 1743
- Effective model calibration via sensible variable identification and adjustment with application to composite fuselage simulation YAN WANG, XIAOWEI YUE, RUI TUO, JEFFREY H. HUNT AND JIANJUN SHI 1759
- A nearest-neighbor based nonparametric test for viral remodeling in heterogeneous single-cell proteomic data TRAMBAK BANERJEE, BHASWAR B. BHATTACHARYA AND GOURAB MUKHERJEE 1777
- The statistical performance of matching-adjusted indirect comparisons: Estimating treatment effects with aggregate external control data
DAVID CHENG, RAJEEV AYYAGARI AND JAMES SIGNOROVITCH 1806
- Mixture of hidden Markov models for accelerometer data
MARIE DU ROY DE CHAUMARAY, MATTHIEU MARBAC AND FABIEN NAVARRO 1834
- RNDCClone: Tumor subclone reconstruction based on integrating DNA and RNA sequence data TIANJIAN ZHOU, SUBHAJIT SENGUPTA, PETER MÜLLER AND YUAN JI 1856
- A Bayesian time-varying effect model for behavioral mHealth data
MATTHEW D. KOSLOVSKY, EMILY T. HÉBERT, MICHAEL S. BUSINELLE AND MARINA VANNUCCI 1878
- Bayesian profiling multiple imputation for missing hemoglobin values in electronic health records YAJUAN SI, MARI PALTA AND MAUREEN SMITH 1903
- Bayesian inference for multistrain epidemics with application to *Escherichia coli* O157:H7 in feedlot cattle PANAYIOTA TOULOUPOU, BÄRBEL FINKENSTÄDT, THOMAS E. BESSER, NIGEL P. FRENCH AND SIMON E. F. SPENCER 1925

continued

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—Continued from front cover

- Integrative statistical methods for exposure mixtures and health
BRIAN J. REICH, YAWEN GUAN, DENIS FOURCHES, JOSHUA L. WARREN,
STEFANIE E. SARNAT AND HOWARD H. CHANG 1945
- Incorporating compositional heterogeneity into Lie Markov models for phylogenetic
inference NAOMI E. HANNAFORD, SARAH E. HEAPS, TOM M. W. NYE,
TOM A. WILLIAMS AND T. MARTIN EMBLEY 1964
- Hawkes binomial topic model with applications to coupled conflict-Twitter data
GEORGE MOHLER, ERIN MCGRATH, CODY BUNTAIN AND GARY LAFREE 1984
- Nonparametric Bayesian multiarmed bandits for single-cell experiment design
FEDERICO CAMERLENGHI, BIANCA DUMITRASCU, FEDERICO FERRARI,
BARBARA E. ENGELHARDT AND STEFANO FAVARO 2003
- Structured discrepancy in Bayesian model calibration for ChemCam on the Mars
Curiosity rover K. SHAM BHAT, KARY MYERS, EARL LAWRENCE,
JAMES COLGAN AND ELIZABETH JUDGE 2020
- Modelling the sound production of narwhals using a point process framework with
memory effects ALEKSANDER SØLTOFT-JENSEN,
MADS PETER HEIDE-JØRGENSEN AND SUSANNE DITLEVSEN 2037
- Region-referenced spectral power dynamics of EEG signals: A hierarchical modeling
approach QIAN LI, JOHN SHAMSHOIAN,
DAMLA ŞENTÜRK, CATHERINE SUGAR, SHAFALI JESTE,
CHARLOTTE DiSTEFANO AND DONATELLO TELESCA 2053
- Mitigating unobserved spatial confounding when estimating the effect of supermarket
access on cardiovascular disease deaths
PATRICK M. SCHNELL AND GEORGIA PAPADOGEORGOU 2069
- Letter to the Editor: Fitting a folded normal distribution without EM
IAIN L. MACDONALD 2096
- Response to ‘Fitting a folded normal distribution without EM’
SUNGKYU JUNG, MARK FOSKEY AND J. S. MARRON 2099
- Rejoinder: Fitting a folded normal distribution without EM IAIN L. MACDONALD 2101

THE ANNALS OF APPLIED STATISTICS

Vol. 14, No. 4, pp. 1581–2101 December 2020

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. *Editor:* Amir Dembo, Department of Statistics and Department of Mathematics, Stanford University, Stanford, California 94305, USA

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 14, Number 4, December 2020. 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, 9650 Rockville Pike, Suite L 2310, Bethesda, Maryland 20814-3998, USA.

HIERARCHICAL MULTIDIMENSIONAL SCALING FOR THE COMPARISON OF MUSICAL PERFORMANCE STYLES

BY ANNA K. YANCHENKO^{*} AND PETER D. HOFF[†]

Department of Statistical Science, Duke University, ^{*}anna.yanchenko@duke.edu; [†]peter.hoff@duke.edu

Quantification of stylistic differences between musical artists is of academic interest to the music community and is also useful for other applications, such as music information retrieval and recommendation systems. Information about stylistic differences can be obtained by comparing the performances of different artists across common musical pieces. In this article we develop a statistical methodology for identifying and quantifying systematic stylistic differences among artists that are consistent across audio recordings of a common set of pieces, in terms of several musical features. Our focus is on a comparison of 10 different orchestras, based on data from audio recordings of the nine Beethoven symphonies. As generative or fully parametric models of raw audio data can be highly complex and more complex than necessary for our goal of identifying differences between orchestras, we propose to reduce the data from a set of audio recordings down to pairwise distances between orchestras, based on different musical characteristics of the recordings, such as tempo, dynamics and timbre. For each of these characteristics, we obtain multiple pairwise distance matrices, one for each movement of each symphony. We develop a hierarchical multidimensional scaling (HMDS) model to identify and quantify systematic differences between orchestras in terms of these three musical characteristics and interpret the results in the context of known qualitative information about the orchestras. This methodology is able to recover several expected systematic similarities between orchestras as well as to identify some more novel results. For example, we find that modern recordings exhibit a high degree of similarity to each other, as compared to older recordings.

REFERENCES

- ANDERSON, M. J. (2001). A new method for non-parametric multivariate analysis. *Austral Ecology* **26** 32–46.
- BAKKER, R. and POOLE, K. T. (2013). Bayesian metric multidimensional scaling. *Polit. Anal.* **21** 125–140.
- BORG, I., GROENEN, P. J. F. and MAIR, P. (2018). *Applied Multidimensional Scaling and Unfolding*. 2nd ed. *SpringerBriefs in Statistics*. Springer, Cham. MR3793129 <https://doi.org/10.1007/978-3-319-73471-2>
- CARROLL, J. D. and CHANG, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart–Young” decomposition. *Psychometrika* **35** 283–319.
- COOK, N. (2005). Towards the complete musicologist? Invited Talk at the International Symposium on Music Information Retrieval.
- DESAIN, P. and HONING, H. (1994). Does expressive timing in music performance scale proportionally with tempo? *Psychol. Res.* **56** 285–292.
- ELLIS, D. (2007). Chroma feature analysis and synthesis. <https://labrosa.ee.columbia.edu/matlab/chroma-ansyn/>.
- FONG, D. K. H., DESARBO, W. S., CHEN, Z. and XU, Z. (2015). A Bayesian vector multidimensional scaling procedure incorporating dimension reparameterization with variable selection. *Psychometrika* **80** 1043–1065. MR3423037 <https://doi.org/10.1007/s11336-015-9449-x>
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284 <https://doi.org/10.1214/06-BA117A>
- 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
- GOEBL, W. and WIDMER, G. (2009). On the use of computational methods for expressive music performance. In *Modern Methods for Musicology: Prospects, Proposals and Realities* 93–113. Ashgate, London.

- GRACHTEN, M., GASSER, M., ARZT, A. and WIDMER, G. (2013). Automatic alignment of music performances with structural differences. In *Proceedings of the 14th International Society for Music Information Retrieval Conference* 607–612.
- INTERNATIONAL MUSIC SCORE LIBRARY PROJECT (2019). IMSLP: Petrucci Music Library. https://imslp.org/wiki/Category:Beethoven,_Ludwig_van.
- KAMMERL, J., BIRKBECK, N., INGUVA, S., KELLY, D., CRAWFORD, A. J., DENMAN, H., KOKARAM, A. and PANTOFARU, C. (2014). Temporal synchronization of multiple audio signals. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 4603–4607.
- KIRCHHOFF, H. and LERCH, A. (2011). Evaluation of features for audio-to-audio alignment. *J. New Music Res.* **40** 27–41.
- KRUSKAL, J. B. (1983). An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM Rev.* **25** 201–237. MR0700350 <https://doi.org/10.1137/1025045>
- KUNSTDERFUGE. COM (2018). Franz Liszt transcriptions. <http://www.kunstderfuge.com/liszt.htm#Transcriptions>.
- LIEBMAN, E., ORNOY, E. and CHOR, B. (2012). A phylogenetic approach to music performance analysis. *J. New Music Res.* **41** 195–222.
- LIEM, C. C. S. and HANJALIC, A. (2011a). Expressive timing from cross-performance and audio-based alignment patterns: An extended case study. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* 519–524.
- LIEM, C. C. S. and HANJALIC, A. (2011b). Expressivity in musical timing in relation to musical structure and interpretation: A cross-performance, audio-based approach. In *Proceedings of the 42nd International AES Conference on Semantic Audio* 255–264.
- LIEM, C. C. S. and HANJALIC, A. (2015). Comparative analysis of orchestral performance recordings: An image-based approach. In *Proceedings of the 16th International Society for Music Information Retrieval Conference* 302–308.
- LIN, L. and FONG, D. K. H. (2019). Bayesian multidimensional scaling procedure with variable selection. *Comput. Statist. Data Anal.* **129** 1–13. MR3855195 <https://doi.org/10.1016/j.csda.2018.07.007>
- LOGAN, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Symposium on Music Information Retrieval*.
- MCARDLE, B. H. and ANDERSON, M. J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology* **8** 290–297.
- MINAS, C. and MONTANA, G. (2014). Distance-based analysis of variance: Approximate inference. *Stat. Anal. Data Min.* **7** 450–470. MR3303162 <https://doi.org/10.1002/sam.11227>
- MÜLLER, M. (2015). *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, Cham. MR3382223 <https://doi.org/10.1007/978-3-319-21945-5>
- OH, M.-S. and RAFTERY, A. E. (2001). Bayesian multidimensional scaling and choice of dimension. *J. Amer. Statist. Assoc.* **96** 1031–1044. MR1947251 <https://doi.org/10.1198/016214501753208690>
- PARK, J., DESARBO, W. S. and LIECHTY, J. (2008). A hierarchical Bayesian multidimensional scaling methodology for accommodating both structural and preference heterogeneity. *Psychometrika* **73** 451–472. MR2447325 <https://doi.org/10.1007/s11336-008-9064-1>
- PENEL, A. and DRAKE, C. (1998). Sources of timing variations in music performance: A psychological segmentation model. *Psychol. Res.* **61** 12–32.
- PEPERKAMP, J., HILDEBRANDT, K. and LIEM, C. C. S. (2017). A formalization of relative local tempo variations in collections of performances. In *Proceedings of the 18th International Society for Music Information Retrieval Conference* 158–164.
- RIZZO, M. L. and SZÉKELY, G. J. (2010). DISCO analysis: A nonparametric extension of analysis of variance. *Ann. Appl. Stat.* **4** 1034–1055. MR2758432 <https://doi.org/10.1214/09-AOAS245>
- SAPP, C. S. (2007). Comparative analysis of multiple musical performances. In *Proceedings of the 8th International Conference on Music Information Retrieval* 497–500.
- SAPP, C. S. (2008). Hybrid numeric/rank similarity metrics for musical performance analysis. In *Proceedings of the 9th International Conference on Music Information Retrieval* 501–506.
- STAN DEVELOPMENT TEAM (2019a). RStan: The R interface to Stan. R package version 2.19.2.
- STAN DEVELOPMENT TEAM (2019b). Stan Reference Manual, 2.21 ed.
- SUEUR, J. (2018). *Sound Analysis and Synthesis with R*. Springer, Berlin.
- SUEUR, J., AUBIN, T., SIMONIS, C., LELLOUCH, L., BROWN, E. C. et al. (2018). Package ‘seewave’: Sound Analysis and Synthesis. <https://cran.r-project.org/web/packages/seewave/seewave.pdf>.
- THORNBURG, H. D., LEISTIKOW, R. J. and BERGER, J. (2007). Melody extraction and musical onset detection from framewise STFT peak data. *IEEE Trans. Audio Speech Lang. Process.* **15** 1257–1272.
- TORGERSON, W. S. (1952). Multidimensional scaling. I. Theory and method. *Psychometrika* **17** 401–419. MR0054219 <https://doi.org/10.1007/BF02288916>

- VAN DEN OORD, A., DIELEMAN, S. and SCHRAUWEN, B. (2013). Deep content-based music recommendation. In *Advances in Neural Information Processing Systems* 26 (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 2643–2651. Curran Associates, Red Hook.
- VAN DEN OORD, A., DIELEMAN, S., ZEN, H., SIMONYAN, K., VINYALS, O., GRAVES, A., KALCHBRENNER, N., SENIOR, A. W. and KAVUKCUOGLU, K. (2016). WaveNet: A generative model for raw audio. [CoRR abs/1609.03499](https://arxiv.org/abs/1609.03499).
- VAN DER MAATEN, L. and HINTON, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** 2579–2605.
- VSEVOLOZHSKAYA, O. A., ZAYKIN, D. V., GREENWOOD, M. C., WEI, C. and LU, Q. (2014). Functional analysis of variance for association studies. *PLoS ONE* **9** 1–13. <https://doi.org/10.1371/journal.pone.0105074>
- YANCHENKO, A. K. and HOFF, P. D. (2020). Supplement to “Hierarchical Multidimensional Scaling for the Comparison of Musical Performance Styles.” <https://doi.org/10.1214/20-AOAS1391SUPPA>, <https://doi.org/10.1214/20-AOAS1391SUPPB>

MEASURING TIMELINESS OF ANNUAL REPORTS FILING BY JUMP ADDITIVE MODELS

BY YICHENG KANG

Department of Mathematical Sciences, Bentley University, ykang@bentley.edu

Foreign public issuers (FPIs) are required by the Securities and Exchanges Commission (SEC) to file Form 20-F as comprehensive annual reports. In an effort to increase the usefulness of 20-Fs, the SEC recently enacted a regulation to accelerate the deadline of 20-F filing from six months to four months after the fiscal year-end. The rationale is that the shortened reporting lag would improve the informational relevance of 20-Fs. In this work we propose a jump additive model to evaluate the SEC's rationale by investigating the relationship between the timeliness of 20-F filing and its decision usefulness using the market data. The proposed model extends the conventional additive models to allow possible discontinuities in the regression functions. We suggest a two-step jump-preserving estimation procedure and show that it is statistically consistent. By applying the procedure to the 20-F study, we find a moderate positive association between the magnitude of the market reaction and the filing timeliness when the acceleration is less than 17 days. We also find that the market considers the filings significantly more informative when the acceleration is more than 18 days and such reaction tapers off when the acceleration exceeds 40 days.

REFERENCES

- ANSELONE, P. M. and LAURENT, P. J. (1968). A general method for the construction of interpolating or smoothing spline-functions. *Numer. Math.* **12** 66–82. [MR0249904 https://doi.org/10.1007/BF02170998](https://doi.org/10.1007/BF02170998)
- ASTHANA, S., BALSAM, S. and SANKARAGURUSWAMY, S. (2004). Differential response of small versus large investors to 10-K filings on EDGAR. *Account. Rev.* **79** 571–589.
- ATIASE, R. K. (1985). Predisclosure information, firm capitalization, and security price behavior around earnings announcements. *J. Acc. Res.* 21–36.
- AVANESOV, V. and BUZUN, N. (2018). Change-point detection in high-dimensional covariance structure. *Electron. J. Stat.* **12** 3254–3294. [MR3861282 https://doi.org/10.1214/18-EJS1484](https://doi.org/10.1214/18-EJS1484)
- BEAVER, W. H. (1968). The information content of annual earnings announcements. *J. Acc. Res.* 67–92.
- BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619. [MR0803258](https://doi.org/10.1080/01621459.1985.10477358)
- BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–555. [MR0994249 https://doi.org/10.1214/aos/1176347115](https://doi.org/10.1214/aos/1176347115)
- CALLEN, J. L., LIVNAT, J. and SEGAL, D. (2006). The information content of SEC filings and information environment: A variance decomposition analysis. *Account. Rev.* **81** 1017–1043.
- CASINI, A. (2018). Tests for forecast instability and forecast failure under a continuous record asymptotic framework. Technical report. Available at [arXiv:1803.10883](https://arxiv.org/abs/1803.10883).
- CASINI, A. (2019). Theory of Evolutionary spectra for heteroskedasticity and autocorrelation robust inference in possibly misspecified and nonstationary models. Technical report, Dept. Economics and Finance, Univ. Rome Tor Vergata.
- CASINI, A. and PERRON, P. (2019). Structural breaks in time series. *Oxford Research Encyclopedia of Economics and Finance*. <https://doi.org/10.1093/acrefore/9780190625979.013.179>
- CHEN, L. H. and SAMI, H. (2008). Trading volume reaction to the earnings reconciliation from IAS to US GAAP. *Contemp. Account. Res.* **25** 15–53.
- CHEN, L. H. and SAMI, H. (2013). The impact of firm characteristics on trading volume reaction to the earnings reconciliation from IFRS to US GAAP. *Contemp. Account. Res.* **30** 697–718.

Key words and phrases. Backfitting, decision usefulness, jump detection, jump-preserving estimation, market reaction, shortened reporting lag.

- CHENG, K. F. and LIN, P. E. (1981). Nonparametric estimation of a regression function. *Z. Wahrsch. Verw. Gebiete* **57** 223–233. [MR0626817](#) <https://doi.org/10.1007/BF00535491>
- CHO, H. and FRYZLEWICZ, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 475–507. [MR3310536](#) <https://doi.org/10.1111/rssb.12079>
- CHRISTENSEN, T. E., HENINGER, W. G. and STICE, E. K. (2013). Factors associated with price reactions and analysts' forecast revisions around SEC filings. *Res. Account. Regul.* **25** 133–148.
- DEMMLER, A. and REINSCH, C. (1975). Oscillation matrices with spline smoothing. *Numer. Math.* **24** 375–382. [MR0395161](#) <https://doi.org/10.1007/BF01437406>
- DE FRANCO, G., WONG, M. F. and ZHOU, Y. (2011). Accounting adjustments and the valuation of financial statement note information in 10-k filings. *Account. Rev.* **86** 1577–1604.
- DOYLE, J. T. and MAGILKE, M. J. (2013). Decision usefulness and accelerated filing deadlines. *J. Acc. Res.* **51** 549–581.
- ETTER, E. R., REES, L. and LUKAWITZ, J. M. (1999). The usefulness to individual and institutional investors of annual earnings announcements and SEC filings by non-US companies. *J. Int. Account. Audit. Tax.* **8** 109–131.
- FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. CRC Press, London. [MR1383587](#)
- FREEMAN, R. N. (1987). The association between accounting earnings and security returns for large and small firms. *J. Account. Econ.* **9** 195–228.
- FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823. [MR0650892](#)
- GASSER, T., KNEIP, A. and KÖHLER, W. (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.* **86** 643–652. [MR1147088](#)
- GASSER, T. and MÜLLER, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation (Proc. Workshop, Heidelberg, 1979)*. *Lecture Notes in Math.* **757** 23–68. Springer, Berlin. [MR0564251](#)
- GIJBELS, I. and GODERNIAUX, A.-C. (2004). Bandwidth selection for changepoint estimation in nonparametric regression. *Technometrics* **46** 76–86. [MR2043389](#) <https://doi.org/10.1198/004017004000000130>
- GIJBELS, I., HALL, P. and KNEIP, A. (2004). Interval and band estimation for curves with jumps *J. Appl. Probab.* **41A** 65–79. [MR2057566](#) <https://doi.org/10.1239/jap/1082552191>
- GRIFFIN, P. A. (2003). Got information? Investor response to form 10-K and form 10-Q EDGAR filings. *Rev. Acc. Stud.* **8** 433–460.
- HALL, P. and TITTERINGTON, D. M. (1992). Edge-preserving and peak-preserving smoothing. *Technometrics* **34** 429–440. [MR1190262](#) <https://doi.org/10.2307/1268942>
- HASTIE, T. and LOADER, C. (1993). Local regression: Automatic kernel carpentry. *Statist. Sci.* **8** 120–129.
- HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. *Monographs on Statistics and Applied Probability* **43**. CRC Press, London. [MR1082147](#)
- 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>
- KANG, Y. (2020a). Consistent blind image deblurring using jump-preserving extrapolation. *J. Comput. Graph. Statist.* **29** 372–382. [MR4116049](#) <https://doi.org/10.1080/10618600.2019.1665536>
- KANG, Y. (2020). Supplement to “Measuring timeliness of annual reports filing by jump additive models.” <https://doi.org/10.1214/20-AOAS1365SUPPA>, <https://doi.org/10.1214/20-AOAS1365SUPPB>
- KANG, Y., GONG, X., GAO, J. and QIU, P. (2019). Errors-in-variables jump regression using local clustering. *Stat. Med.* **38** 3642–3655. [MR3999222](#) <https://doi.org/10.1002/sim.8205>
- KIM, Y., LI, H. and LI, S. (2012). Does eliminating the form 20-F reconciliation from IFRS to US GAAP have capital market consequences? *J. Account. Econ.* **53** 249–270.
- KIM, O. and VERRECCHIA, R. E. (1991). Trading volume and price reactions to public announcements. *J. Acc. Res.* 302–321.
- 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>
- LIU, Z. (2016). Effect of Shortened Reporting Lag on the Usefulness of Form 20-F Ph.D. thesis. Florida International Univ.
- LOADER, C. R. (1999). Bandwidth selection: Classical or plug-in? *Ann. Statist.* **27** 415–438. [MR1714723](#) <https://doi.org/10.1214/aos/1018031201>
- MA, S. and YANG, L. (2011). A jump-detecting procedure based on spline estimation. *J. Nonparametr. Stat.* **23** 67–81. [MR2780816](#) <https://doi.org/10.1080/10485250903571978>

- MEEK, G. K. (1983). US securities market responses to alternate earnings disclosures of non-US multinational corporations. *Account. Rev.* 394–402.
- MÜLLER, C. H. (2002). Robust estimators for estimating discontinuous functions *Metrika* **55** 99–109. [MR1903286 https://doi.org/10.1007/s001840200190](https://doi.org/10.1007/s001840200190)
- NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.
- OLIBE, K. O. (2001). Assessing the usefulness of SEC form 20-F disclosures using return and volume metrics: The case of UK firms. *J. Econ. Finance* **25** 343–357.
- OPSOMER, J. D. and RUPPERT, D. (1998). A fully automated bandwidth selection method for fitting additive models. *J. Amer. Statist. Assoc.* **93** 605–619. [MR1631333 https://doi.org/10.2307/2670112](https://doi.org/10.2307/2670112)
- PRIESTLEY, M. B. and CHAO, M. T. (1972). Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34** 385–392. [MR0331616](https://doi.org/10.2307/2334166)
- QIU, P. (2005). *Image Processing and Jump Regression Analysis*. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. [MR2111430 https://doi.org/10.1002/0471733156](https://doi.org/10.1002/0471733156)
- ROUSSEEUW, P. J. and CROUX, C. (1993). Alternatives to the median absolute deviation. *J. Amer. Statist. Assoc.* **88** 1273–1283. [MR1245360](https://doi.org/10.2307/2286683)
- RUPPERT, D., SHEATHER, S. J. and WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 1257–1270. [MR1379468](https://doi.org/10.2307/2286683)
- WAHBA, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. [MR1045442 https://doi.org/10.1137/1.9781611970128](https://doi.org/10.1137/1.9781611970128)
- WANG, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika* **82** 385–397. [MR1354236 https://doi.org/10.1093/biomet/82.2.385](https://doi.org/10.1093/biomet/82.2.385)
- WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372. [MR0185765](https://doi.org/10.2307/2334166)
- WU, W. B. and ZHAO, Z. (2007). Inference of trends in time series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 391–410. [MR2323759 https://doi.org/10.1111/j.1467-9868.2007.00594.x](https://doi.org/10.1111/j.1467-9868.2007.00594.x)
- XIA, Z. and QIU, P. (2015). Jump information criterion for statistical inference in estimating discontinuous curves. *Biometrika* **102** 397–408. [MR3371012 https://doi.org/10.1093/biomet/asv018](https://doi.org/10.1093/biomet/asv018)
- YOU, H. and ZHANG, X.-J. (2009). Financial reporting complexity and investor underreaction to 10-K information. *Rev. Acc. Stud.* **14** 559–586.

MONOTONIC EFFECTS OF CHARACTERISTICS ON RETURNS

BY JARED D. FISHER¹, DAVID W. PUELZ² AND CARLOS M. CARVALHO³

¹*Department of Statistics, University of California, Berkeley, jared.fisher@berkeley.edu*

²*Booth School of Business, University of Chicago, david.puelz@chicagobooth.edu*

³*McCombs School of Business, University of Texas at Austin, carlos.carvalho@mcombs.utexas.edu*

This paper considers the problem of modeling a firm's expected return as a nonlinear function of its observable characteristics. We investigate whether theoretically-motivated monotonicity constraints on characteristics and non-stationarity of the conditional expectation function provide statistical and economic benefit. We present an interpretable model that has similar out-of-sample performance to black-box machine learning methods. With this model, the data provide support for monotonicity and time variability of the conditional expectation function. Additionally, we develop an approach for characteristic selection using loss functions to summarize the posterior distribution. Standard unexplained volume, short-term reversal, size, and variants of momentum are found to be significant characteristics, and there is evidence this set changes over time.

REFERENCES

- ANG, A. and KRISTENSEN, D. (2012). Testing conditional factor models. *J. Financ. Econ.* **106** 132–156. <https://doi.org/10.1016/j.jfineco.2012.04.008>
- ANG, A., HODRICK, R., XING, Y. and ZHANG, X. (2006). The cross-section of volatility and expected returns. *J. Finance* **61** 259–299. <https://doi.org/10.1111/j.1540-6261.2006.00836.x>
- BALAKRISHNAN, K., BARTOV, E. and FAUREL, L. (2010). Post loss/profit announcement drift. *J. Account. Econ.* **50** 20–41. <https://doi.org/10.1016/j.jacceco.2009.12.002>
- BALL, R., GERAKOS, J., LINNAINMAA, J. T. and NIKOLAEV, V. V. (2015). Deflating profitability. *J. Financ. Econ.* **117** 225–248. <https://doi.org/10.1016/j.jfineco.2015.02.004>
- BASU, S. (1983). The relationship between earnings' yield, market value and return for NYSE common stocks: Further evidence. *J. Financ. Econ.* **12** 129–156. [https://doi.org/10.1016/0304-405X\(83\)90031-4](https://doi.org/10.1016/0304-405X(83)90031-4)
- BEKAERT, G., HARVEY, C. R. and LUMSDAINE, R. L. (2002). Dating the integration of world equity markets. *J. Financ. Econ.* **65** 203–247. [https://doi.org/10.1016/S0304-405X\(02\)00139-3](https://doi.org/10.1016/S0304-405X(02)00139-3)
- BHANDARI, C. L. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *J. Finance* **43** 507–528. <https://doi.org/10.1111/j.1540-6261.1988.tb03952.x>
- BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.
- BUSTAMANTE, C. M. and DONANGELO, A. (2017). Product market competition and industry returns. *Rev. Financ. Stud.* **30** 4216–4266. <https://doi.org/10.1093/rfs/hhx033>
- CATTANEO, M. D., CRUMP, R. K., FARRELL, M. H. and SCHAUMBURG, E. (2020). Characteristic-sorted portfolios: Estimation and inference. *Rev. Econ. Stat.* **102** 531–551. https://doi.org/10.1162/rest_a_00883
- CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. <https://doi.org/10.1214/09-AOAS285>
- CHIPMAN, H. A., GEORGE, E. I., MCCULLOCH, R. E. and SHIVELY, T. S. (2019). High-dimensional nonparametric monotone function estimation using BART. Preprint. Available at [arXiv:1612.01619](https://arxiv.org/abs/1612.01619).
- CHUNG, K. H. and ZHANG, H. (2014). A simple approximation of intraday spreads using daily data. *J. Financ. Mark.* **17** 94–120. <https://doi.org/10.1016/j.finmar.2013.02.004>
- COCHRANE, J. H. (2011). Presidential address: Discount rates. *J. Finance* **66** 1047–1108. <https://doi.org/10.1111/j.1540-6261.2011.01671.x>
- COOPER, M., GULEN, H. and SCHILL, M. (2008). Asset growth and the cross-section of stock returns. *J. Finance* **63** 1609–1651. <https://doi.org/10.1111/j.1540-6261.2008.01370.x>
- D'ACUNTO, F., LIU, R., PFLUEGER, C. and WEBER, M. (2018). Flexible prices and leverage. *J. Financ. Econ.* **129** 46–68. <https://doi.org/10.1016/j.jfineco.2018.03.009>

- DATAR, V. T., NAIK, N. Y. and RADCLIFFE, R. (1998). Liquidity and stock returns: An alternative test. *J. Financ. Mark.* **1** 203–219. [https://doi.org/10.1016/S1386-4181\(97\)00004-9](https://doi.org/10.1016/S1386-4181(97)00004-9)
- DE BONDT, W. and THALER, R. (1985). Does the stock market overreact? *J. Finance* **40** 793–805. <https://doi.org/10.1111/j.1540-6261.1985.tb05004.x>
- FAMA, E. and FRENCH, K. (1992). The cross-section of expected stock returns. *J. Finance* **47** 427–465. <https://doi.org/10.1111/j.1540-6261.1992.tb04398.x>
- FAMA, E. F. and FRENCH, K. R. (1993). Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* **33** 3–56. [https://doi.org/10.1016/0304-405X\(93\)90023-5](https://doi.org/10.1016/0304-405X(93)90023-5)
- FAMA, E. F. and FRENCH, K. R. (2008). Dissecting anomalies. *J. Finance* **63** 1653–1678. <https://doi.org/10.1111/j.1540-6261.2008.01371.x>
- FAMA, E. F. and FRENCH, K. R. (2016). Dissecting anomalies with a five-factor model. *Rev. Financ. Stud.* **29** 69–103. <https://doi.org/10.1093/rfs/hhv043>
- FAMA, E. F. and MACBETH, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *J. Polit. Econ.* **81** 607–636.
- FREYBERGER, J., NEUHIERL, A. and WEBER, M. (2020). Dissecting characteristics nonparametrically. *Rev. Financ. Stud.* **33** 2326–2377. <https://doi.org/10.1093/rfs/hhz123>
- GANDHI, P. and LUSTIG, H. (2015). Size anomalies in U.S. bank stock returns. *J. Finance* **70** 733–768. <https://doi.org/10.1111/jofi.12235>
- GARFINKEL, J. (2009). Measuring investors' opinion divergence. *J. Acc. Res.* **47** 1317–1348. <https://doi.org/10.1111/j.1475-679X.2009.00344.x>
- GEORGE, T. and HWANG, C. (2004). The 52-week high and momentum investing. *J. Finance* **59** 2145–2176. <https://doi.org/10.1111/j.1540-6261.2004.00695.x>
- GORODNICHENKO, Y. and WEBER, M. (2016). Are sticky prices costly? Evidence from the stock market. *Am. Econ. Rev.* **106** 165–99. <https://doi.org/10.1257/aer.20131513>
- GU, S., KELLY, B. and XIU, D. (2020). Empirical asset pricing via machine learning. *Rev. Financ. Stud.* <https://doi.org/10.1093/rfs/hhaa009>
- HAHN, P. R. and CARVALHO, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *J. Amer. Statist. Assoc.* **110** 435–448. <https://doi.org/10.1080/01621459.2014.993077>
- HAN, Y., HE, A., RAPACH, D. and ZHOU, G. (2019). Firm characteristics and expected stock returns. Working paper.
- HARVEY, C. R., LIU, Y. and ZHU, H. (2016). ... and the cross-section of expected returns. *Rev. Financ. Stud.* **29** 5–68. <https://doi.org/10.1093/rfs/hhv059>
- HAUGEN, R. A. and BAKER, N. L. (1996). Commonality in the determinants of expected stock returns. *J. Financ. Econ.* **41** 401–439. [https://doi.org/10.1016/0304-405X\(95\)00868-F](https://doi.org/10.1016/0304-405X(95)00868-F)
- HE, J., YALOV, S. and HAHN, P. R. (2019). XBART: Accelerated Bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- HIRSHLEIFER, D., KEWEI, H., TEOH, S. H. and YINGLEI, Z. (2004). Do investors overvalue firms with bloated balance sheets? *J. Account. Econ.* **38** 297–331. <https://doi.org/10.1016/j.jacceco.2004.10.002>
- HOU, K., KAROLYI, A. G. and KHO, B.-C. (2011). What Factors Drive Global Stock Returns?. *Rev. Financ. Stud.* **24** 2527–2574.
- JEGADEESH, N. and TITMAN, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *J. Finance* **48** 65–91.
- JEGADEESH, N. and TITMAN, S. (2001). Profitability of momentum strategies: An evaluation of alternative explanations. *J. Finance* **56** 699–720.
- LETTAU, M. and NIEUWERBURGH, S. V. (2008). Reconciling the return predictability evidence. *Rev. Financ. Stud.* **21** 1607–1652.
- LEWELLEN, J. (2015). The cross-section of expected stock returns. *Crit. Finance Rev.* **4** 1–44. <https://doi.org/10.1561/104.000000024>
- LYANDRES, E., SUN, L. and ZHANG, L. (2007). The new issues puzzle: Testing the investment-based explanation. *Rev. Financ. Stud.* **21** 2825–2855. <https://doi.org/10.1093/rfs/hhm058>
- MCCARTHY, D. and JENSEN, S. T. (2016). Power-weighted densities for time series data. *Ann. Appl. Stat.* **10** 305–334. <https://doi.org/10.1214/15-AOAS893>
- NOVY-MARX, R. (2010). Operating leverage. *Rev. Finance* **15** 103–134. <https://doi.org/10.1093/rof/rfq019>
- NOVY-MARX, R. (2012). Is momentum really momentum? *J. Financ. Econ.* **103** 429–453. <https://doi.org/10.1016/j.jfineco.2011.05.003>
- PALAZZO, B. (2012). Cash holdings, risk, and expected returns. *J. Financ. Econ.* **104** 162–185. <https://doi.org/10.1016/j.jfineco.2011.12.009>
- PASTOR, L. and STAMBAUGH, R. F. (2001). The equity premium and structural breaks. *J. Finance* **56** 1207–1239. <https://doi.org/10.1111/0022-1082.00365>

- PATTON, A. J. and TIMMERMANN, A. (2010). Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and portfolio sorts. *J. Financ. Econ.* **98** 605–625.
- PETTENUZZO, D. and TIMMERMANN, A. (2011). Predictability of stock returns and asset allocation under structural breaks. *J. Econometrics* **164** 60–78. [MR2821794](#) <https://doi.org/10.1016/j.jeconom.2011.02.019>
- PUELZ, D. W. (2018). Regularization in econometrics and finance. Ph.D. thesis, Univ. Texas at Austin, McCombs School of Business.
- PUELZ, D., CARVALHO, C. M. and HAHN, P. R. (2015). Optimal ETF selection for passive investing.
- PUELZ, D., HAHN, P. R. and CARVALHO, C. M. (2017). Variable selection in seemingly unrelated regressions with random predictors. *Bayesian Anal.* **12** 969–989. [MR3724975](#) <https://doi.org/10.1214/17-BA1053>
- PUELZ, D., HAHN, P. R. and CARVALHO, C. M. (2020). Portfolio selection for individual passive investing. *Appl. Stoch. Models Bus. Ind.* **36** 124–142.
- SHIVELY, T. S., SAGER, T. W. and WALKER, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 159–175. [MR2655528](#) <https://doi.org/10.1111/j.1467-9868.2008.00677.x>
- SLOAN, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *Account. Rev.* **71** 289–315.
- SMITH, S. and TIMMERMANN, A. (2018). Break risk. Working paper.
- SMITH, S. C., TIMMERMANN, A. and ZHU, Y. (2019). Variable selection in panel models with breaks. *J. Econometrics* **212** 323–344. [MR3994020](#) <https://doi.org/10.1016/j.jeconom.2019.04.033>
- SOLIMAN, M. T. (2008). The use of DuPont analysis by market participants. *Account. Rev.* **83** 823–853.
- WRIGHT, M. N. and ZIEGLER, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77** Issue 1. <https://doi.org/10.18637/jss.v077.i01>
- YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. [MR2212574](#) <https://doi.org/10.1111/j.1467-9868.2005.00532.x>

DATA FUSION MODEL FOR SPECIATED NITROGEN TO IDENTIFY ENVIRONMENTAL DRIVERS AND IMPROVE ESTIMATION OF NITROGEN IN LAKES

BY ERIN M. SCHLIEP^{1,*}, SARAH M. COLLINS², SHIRLEY ROJAS-SALAZAR^{1,†},
NOAH R. LOTTIG^{3,‡} AND EMILY H. STANLEY^{3,§}

¹Department of Statistics, University of Missouri, *schliepe@missouri.edu; †srnw3@mail.missouri.edu

²Department of Zoology and Physiology, University of Wyoming, sarah.collins@uwyo.edu

³Center for Limnology, University of Wisconsin, ‡nrlottig@wisc.edu; §ehstanley@wisc.edu

Concentrations of nitrogen provide a critical metric for understanding ecosystem function and water quality in lakes. However, varying approaches for quantifying nitrogen concentrations may bias the comparison of water quality across lakes and regions. Different measurements of total nitrogen exist based on its composition (e.g., organic versus inorganic, dissolved versus particulate), which we refer to as nitrogen species. Fortunately, measurements of multiple nitrogen species are often collected and can, therefore, be leveraged together to inform our understanding of the controls on total nitrogen in lakes. We develop a multivariate hierarchical statistical model that fuses speciated nitrogen measurements, obtained across multiple methods of reporting, in order to improve our estimates of total nitrogen. The model accounts for lower detection limits and measurement error that vary across lake, species and observation. By modeling speciated nitrogen, as opposed to previous efforts that mostly consider only total nitrogen, we obtain more resolved inference with regard to differences in sources of nitrogen and their relationship with complex environmental drivers. We illustrate the inferential benefits of our model using speciated nitrogen data from the LAke GeOSpatial and temporal database (LAGOS).

REFERENCES

- BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *J. Agric. Biol. Environ. Stat.* **15** 176–197. [MR2787270 https://doi.org/10.1007/s13253-009-0004-z](https://doi.org/10.1007/s13253-009-0004-z)
- CAMERON, K., DI, H. J. and MOIR, J. (2013). Nitrogen losses from the soil/plant system: A review. *Ann. Appl. Biol.* **162** 145–173.
- CHEN, M., ZENG, G., ZHANG, J., XU, P., CHEN, A. and LU, L. (2015). Global landscape of total organic carbon, nitrogen and phosphorus in lake water. *Sci. Rep.* **5** 15043. <https://doi.org/10.1038/srep15043>
- 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.
- COLLINS, S. M., OLIVER, S. K., LAPIERRE, J.-F., STANLEY, E. H., JONES, J. R., WAGNER, T. and SORANNO, P. A. (2017). Lake nutrient stoichiometry is less predictable than nutrient concentrations at regional and sub-continental scales. *Ecol. Appl.* **27** 1529–1540. <https://doi.org/10.1002/eap.1545>
- DORIOZ, J. and FERHI, A. (1994). Non-point pollution and management of agricultural areas: Phosphorus and nitrogen transfer in an agricultural watershed. *Water Res.* **28** 395–410.
- FERGUS, C. E., LAPIERRE, J., OLIVER, S. K., SKAFF, N. K., CHERUVELIL, K. S., CAREN SCOTT, K. E. W. and SORANNO, P. A. (2017). The freshwater landscape: Lake, wetland, and stream abundance and connectivity at macroscales. *Ecosphere* **8** e01911.
- FUENTES, M. and RAFTERY, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61** 36–45. [MR2129199 https://doi.org/10.1111/j.0006-341X.2005.030821.x](https://doi.org/10.1111/j.0006-341X.2005.030821.x)

Key words and phrases. Bayesian hierarchical model, detection limits, LAGOS, multivariate, Markov chain Monte Carlo.

- GLIBERT, P. M., WILKERSON, F. P., DUGDALE, R. C., RAVEN, J. A., DUPONT, C. L., LEAVITT, P. R., PARKER, A. E., BURKHOLDER, J. M. and KANA, T. M. (2016). Pluses and minuses of ammonium and nitrate uptake and assimilation by phytoplankton and implications for productivity and community composition, with emphasis on nitrogen-enriched conditions. *Limnol. Oceanogr.* **61** 165–197.
- GOBLER, C. J., BURKHOLDER, J. M., DAVIS, T. W., HARKE, M. J., JOHNGEN, T., STOW, C. A. and DE WAAL, D. B. V. (2016). The dual role of nitrogen supply in controlling the growth and toxicity of cyanobacterial blooms. *Harmful Algae* **54** 87–97. <https://doi.org/10.1016/j.hal.2016.01.010>
- GUILLAS, S., BAO, J., CHOI, Y. and WANG, Y. (2008). Statistical correction and downscaling of chemical transport model ozone forecasts over Atlanta. *Atmos. Environ.* **42** 1338–1348.
- HARPOLE, W. S., NGAI, J. T., CLELAND, E. E., SEABLOOM, E. W., BORER, E. T., BRACKEN, M. E., ELSER, J. J., GRUNER, D. S., HILLEBRAND, H. et al. (2011). Nutrient co-limitation of primary producer communities. *Ecol. Lett.* **14** 852–862.
- HILKER, T., WULDER, M. A., COOPS, N. C., LINKE, J., McDERMID, G., MASEK, J. G., GAO, F. and WHITE, J. C. (2009). A new data fusion model for high spatial-and temporal-resolution mapping of forest disturbance based on Landsat and MODIS. *Remote Sens. Environ.* **113** 1613–1627.
- HILL, R. A., WEBER, M. H., DEBBOUT, R. M., LEIBOWITZ, S. G. and OLSEN, A. R. (2018). The lake-catchment (LakeCat) dataset: Characterizing landscape features for lake basins within the conterminous USA. *Freshw. Sci.* **37** 208–221. <https://doi.org/10.1086/697966>
- KNOLL, L. B., HAGENBUCH, E. J., STEVENS, M. H., VANNI, M. J., RENWICK, W. H., DENLINGER, J. C., HALE, R. S. and GONZÁLEZ, M. J. (2015). Predicting eutrophication status in reservoirs at large spatial scales using landscape and morphometric variables. *Inland Waters* **5** 203–214.
- OLIVER, S. K., COLLINS, S. M., SORANNO, P. A., WAGNER, T., STANLEY, E. H., JONES, J. R., STOW, C. A. and LOTTIG, N. R. (2017). Unexpected stasis in a changing world: Lake nutrient and chlorophyll trends since 1990. *Glob. Change Biol.* **23** 5455–5467.
- PACIFICI, K., REICH, B. J., MILLER, D. A. W., GARDNER, B., STAUFFER, G., SINGH, S., MCKERROW, A. and COLLAZO, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology* **98** 840–850. <https://doi.org/10.1002/ecy.1710>
- PAERL, H. W., SCOTT, J. T., MCCARTHY, M. J., NEWELL, S. E., GARDNER, W. S., HAVENS, K. E., HOFFMAN, D. K., WILHEIM, S. W. and WURTSBAUGH, W. A. (2016). It takes two to tango: When and where dual nutrient (N & P) reductions are needed to protect lakes and downstream ecosystems. *Environ. Sci. Technol. Lett.* **50** 10805–10813.
- READ, E. K., PATIL, V. P., OLIVER, S. K., HETHERINGTON, A. L., BRENTUP, J. A., ZWART, J. A., WINTERS, K. M., CORMAN, J. R., NODINE, E. R. et al. (2015). The importance of lake-specific characteristics for water quality across the continental United States. *Ecol. Appl.* **25** 943–955.
- RUNDEL, C. W., SCHLIEP, E. M., GELFAND, A. E. and HOLLAND, D. M. (2015). A data fusion approach for spatial analysis of speciated PM_{2.5} across time. *Environmetrics* **26** 515–525. [MR3431926 https://doi.org/10.1002/env.2369](https://doi.org/10.1002/env.2369)
- SAUNDERS, J. F. III, YU, Y., MCCUTCHAN, J. H. JR and ROSARIO-ORTIZ, F. L. (2017). Characterizing limits of precision for dissolved organic nitrogen calculations. *Environ. Sci. Technol. Lett.* **4** 452–456.
- SCHINDLER, D. (2012). The dilemma of controlling cultural eutrophication in lakes. *Proc. R. Soc. Lond., B Biol. Sci.* **279** 4322–4333.
- SCHINDLER, D. W., HECKY, R. E., FINDLAY, D. L., STANTON, M. P., PARKER, B. R., PATERSON, M. J., BEATY, K. G., LYNG, M. and KASIAN, S. E. M. (2008). Eutrophication of lakes cannot be controlled by reducing nitrogen input: Results of a 37-year whole-ecosystem experiment. *Proc. Natl. Acad. Sci. USA* **105** 11254–11258. <https://doi.org/10.1073/pnas.0805108105>
- SCHLIEP, E. M., COLLINS, S. M., ROJAS-SALAZAR, S., LOTTIG, N. R. and STANLEY, E. H. (2020). Supplement to “Data fusion model for speciated nitrogen to identify environmental drivers and improve estimation of nitrogen in lakes.” <https://doi.org/10.1214/20-AOAS1371SUPP>
- SORANNO, P. A., BISSELL, E. G., CHERUVELIL, K. S., CHRISTEL, S. T., COLLINS, S. M., FERGUS, C. E., FILSTRUP, C. T., LAPIERRE, J.-F., LOTTIG, N. R. et al. (2015a). Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science and data reuse. *GigaScience* **4** 28. <https://doi.org/10.1186/s13742-015-0067-4>
- SORANNO, P. A., CHERUVELIL, K. S., WAGNER, T., WEBSTER, K. E. and BREMIGAN, M. T. (2015b). Effects of land use on lake nutrients: The importance of scale, hydrologic connectivity, and region. *PLoS ONE*. **10** e0135454.
- SORANNO, P. A. and CHERUVELIL, K. S. (2017). LAGOS-NE-LIMNO v1.087.1: A module for LAGOS-NE, a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of U.S. lakes: 1925–2013. Environmental Data Initiative.

- SORANNO, P. A., BACON, L. C., BEAUCHENE, M., BEDNAR, K. E., BISSELL, E. G., BOUDREAU, C. K., BOYER, M. G., BREMIGAN, M. T., CARPENTER, S. R. et al. (2017). LAGOS-NE: A multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of US lakes. *GigaScience* **6** 1–22. <https://doi.org/10.1093/gigascience/gix101>
- SORANNO, P. A., WAGNER, T., COLLINS, S. M., LAPIERRE, J.-F., LOTTIG, N. R. and OLIVER, S. K. (2019). Spatial and temporal variation of ecosystem properties at macroscales. *Ecol. Lett.* **22** 1587–1598.
- STACHELEK, J. and OLIVER, S. (2017). LAGOSNE: R interface to the lake multi-scaled geospatial and temporal database. R package version 1.0.0.
- STANLEY, E. H., COLLINS, S. M., LOTTIG, N. R., OLIVER, S. K., WEBSTER, K. E., CHERUVELIL, K. S. and SORANNO, P. A. (2019a). Biases in lake water quality sampling and implications for macroscale research. *Limnol. Oceanogr.* **10** e0135454.
- STANLEY, E. H., SALAZAR, S. R., SCHLIEP, E. M., LOTTIG, N. R., FILSTRUP, C. T. and COLLINS, S. M. (2019b). Comparison of total nitrogen data from direct and kjeldahl-based approaches in integrated datasets. *Limnol. Oceanogr., Methods* **17** 639–649.
- STOW, C. A., WEBSTER, K. E., WAGNER, T., LOTTIG, N., SORANNO, P. A. and CHA, Y. (2018). Small values in big data: The continuing need for appropriate metadata. *Ecol. Inform.* **45** 26–30.
- WAGNER, T. and SCHLIEP, E. M. (2018). Combining nutrient, productivity, and landscape-based regressions improves predictions of lake nutrients and provides insight into nutrient coupling at macroscales. *Limnol. Oceanogr.* **63** 2372–2383.
- WETZEL, R. G. (2001). *Limnology: Lake and River Ecosystems*. Gulf Professional Publishing.
- WU, Z., LIU, Y., LIANG, Z., WU, S. and GUO, H. (2017). Internal cycling, not external loading, decides the nutrient limitation in eutrophic lake: A dynamic model with temporal Bayesian hierarchical inference. *Water Res.* **116** 231–240.

FEATURE SELECTION FOR DATA INTEGRATION WITH MIXED MULTIVIEW DATA

BY YULIA BAKER¹, TIFFANY M. TANG² AND GENEVERA I. ALLEN³

¹Department of Statistics, Rice University, yulia.baker@rice.edu

²Department of Statistics, University of California, tiffany.tang@berkeley.edu

³Department of Electrical and Computer Engineering, Rice University, gallen@rice.edu

Data integration methods that analyze multiple sources of data simultaneously can often provide more holistic insights than can separate inquiries of each data source. Motivated by the advantages of data integration in the era of “big data,” we investigate feature selection for high-dimensional multiview data with mixed data types (e.g., continuous, binary, count-valued). This heterogeneity of multiview data poses numerous challenges for existing feature selection methods. However, after critically examining these issues through empirical and theoretically-guided lenses, we develop a practical solution, the Block Randomized Adaptive Iterative Lasso (B-RAIL) which combines the strengths of the randomized Lasso, adaptive weighting schemes and stability selection. B-RAIL serves as a versatile data integration method for sparse regression and graph selection, and we demonstrate the effectiveness of B-RAIL through extensive simulations and a case study to infer the ovarian cancer gene regulatory network. In this case study, B-RAIL successfully identifies well-known biomarkers associated with ovarian cancer and hints at novel candidates for future ovarian cancer research.

REFERENCES

- ACAR, E., KOLDA, T. G. and DUNLAVY, D. M. (2011). All-at-once optimization for coupled matrix and tensor factorizations. ArXiv Preprint. Available at [arXiv:1105.3422](https://arxiv.org/abs/1105.3422).
- ALLEN, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16** 125–127. [MR0343481 https://doi.org/10.2307/1267500](https://doi.org/10.2307/1267500)
- ANTONIOU, A., PHAROAH, P., NAROD, S., RISCH, H. A., EYFJORD, J. E., HOPPER, J., LOMAN, N., OLSSON, H., JOHANSSON, O. et al. (2003). Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies. *Am. J. Hum. Genet.* **72** 1117–1130.
- BAKER, Y., TANG, T. M and ALLEN, G. I (2020). Supplement to “Feature Selection for Data Integration with Mixed Multiview Data.” <https://doi.org/10.1214/20-AOAS1389SUPP>
- BOREN, T., XIONG, Y., HAKAM, A., WENHAM, R., APTE, S., CHAN, G., KAMATH, S. G., CHEN, D.-T., DRESSMAN, H. et al. (2009). MicroRNAs and their target messenger RNAs associated with ovarian cancer response to chemotherapy. *Gynecol. Oncol.* **113** 249–255.
- BRCA, S. G. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266** 7.
- BUCHHOLTZ, M.-L., BRÜNING, A., MYLONAS, I. and JÜCKSTOCK, J. (2014). Epigenetic silencing of the LDOC1 tumor suppressor gene in ovarian cancer cells. *Arch. Gynecol. Obstet.* **290** 149–154.
- BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. [MR3102549 https://doi.org/10.3150/12-BEJSP11](https://doi.org/10.3150/12-BEJSP11)
- BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. [MR2312149 https://doi.org/10.1214/07-EJS008](https://doi.org/10.1214/07-EJS008)
- CHEN, J. and CHEN, Z. (2012). Extended BIC for small- n -large- P sparse GLM. *Statist. Sinica* **22** 555–574. [MR2954352 https://doi.org/10.5705/ss.2010.216](https://doi.org/10.5705/ss.2010.216)
- CHEN, S., WITTEN, D. M. and SHOJAIE, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102** 47–64. [MR3335095 https://doi.org/10.1093/biomet/asu051](https://doi.org/10.1093/biomet/asu051)

- CHENG, J., LI, T., LEVINA, E. and ZHU, J. (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Statist.* **26** 367–378. [MR3640193](#) <https://doi.org/10.1080/10618600.2016.1237362>
- 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>
- FREIER, C. (2016). Role of regulatory T cells and associated chemokines in human gynecological tumors Ph.D. thesis LMU.
- GAO, J.-Q., TSUDA, Y., HAN, M., XU, D.-H., KANAGAWA, N., HATANAKA, Y., TANI, Y., MIZUGUCHI, H., TSUTSUMI, Y. et al. (2009). NK cells are migrated and indispensable in the anti-tumor activity induced by CCL27 gene therapy. *Cancer Immunology, Immunotherapy* **58** 291.
- GENG, J., LUO, H., PU, Y., ZHOU, Z., WU, X., XU, W. and YANG, Z. (2012). Methylation mediated silencing of miR-23b expression and its role in glioma stem cells. *Neurosci. Lett.* **528** 185–189.
- GIANNAKAKIS, A., SANDALTZOPOULOS, R., GRESHOCK, J., LIANG, S., HUANG, J., HASEGAWA, K., LI, C., O'BRIEN-JENKINS, A., KATSAROS, D. et al. (2008). miR-210 links hypoxia with cell cycle regulation and is deleted in human epithelial ovarian cancer. *Cancer Biol. Ther.* **7** 255–264.
- HALL, D. L. and LLINAS, J. (1997). An introduction to multisensor data fusion. *Proc. IEEE* **85** 6–23.
- HASLBECK, J. and WALDORP, L. J. (2015). mgm: Structure estimation for time-varying mixed graphical models in high-dimensional data. ArXiv Preprint. Available at [arXiv:1510.06871](#).
- JALALI, A., RAVIKUMAR, P., VASUKI, V. and SANGHAVI, S. (2011). On learning discrete graphical models using group-sparse regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 378–387.
- KING, M.-C., MARKS, J. H., MANDELL, J. B. et al. (2003). Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* **302** 643–646.
- LEE, J. and HASTIE, T. (2013). Structure learning of mixed graphical models. In *Artificial Intelligence and Statistics* 388–396.
- LI, W., LIU, Z., CHEN, L., ZHOU, L. and YAO, Y. (2014). MicroRNA-23b is an independent prognostic marker and suppresses ovarian cancer progression by targeting runt-related transcription factor-2. *FEBS Lett.* **588** 1608–1615.
- LIU, H., ROEDER, K. and WASSERMAN, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in Neural Information Processing Systems* 1432–1440.
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. [MR2278363](#) <https://doi.org/10.1214/009053606000000281>
- MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. [MR2758523](#) <https://doi.org/10.1111/j.1467-9868.2010.00740.x>
- MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. [MR2488351](#) <https://doi.org/10.1214/07-AOS582>
- NELSEN, R. B. (1999). *An Introduction to Copulas. Lecture Notes in Statistics* **139**. Springer, New York. [MR1653203](#) <https://doi.org/10.1007/978-1-4757-3076-0>
- OBERMAYR, E., SANCHEZ-CABO, F., TEA, M.-K. M., SINGER, C. F., KRAINER, M., FISCHER, M. B., SEHOULI, J., REINTHALLER, A., HORVAT, R. et al. (2010). Assessment of a six gene panel for the molecular detection of circulating tumor cells in the blood of female cancer patients. *BMC Cancer* **10** 666.
- RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1 -regularized logistic regression. *Ann. Statist.* **38** 1287–1319. [MR2662343](#) <https://doi.org/10.1214/09-AOS691>
- SHAO, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88** 486–494. [MR1224373](#)
- SHEN, R., OLSHEN, A. B. and LADANYI, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25** 2906–2912.
- STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36** 111–147. [MR0356377](#)
- SU, W., BOGDAN, M. and CANDÈS, E. (2017). False discoveries occur early on the Lasso path. *Ann. Statist.* **45** 2133–2150. [MR3718164](#) <https://doi.org/10.1214/16-AOS1521>
- THE CANCER GENOME ATLAS RESEARCH NETWORK (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* **474** 609.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- TIBSHIRANI, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.* **7** 1456–1490. [MR3066375](#) <https://doi.org/10.1214/13-EJS815>
- TONG, M., WONG, T. L., LUK, S. T.-C., CHE, N., WONG, K. Y., FUNG, T. M., GUAN, X.-Y., LEE, N. P., YUAN, Y.-F. et al. (2017). Down-regulation of 4-hydroxyphenylpyruvate dioxygenase (HPD) contributes to the pathogenesis of hepatocellular carcinoma (HCC) through ERK/BCL-2 signalling activation.

- TOYAMA, T., IWASE, H., WATSON, P., MUZIK, H., SAETTLER, E., MAGLIOCCO, A., DiFRANCESCO, L., FORSYTH, P., GARKAVTSEV, I. et al. (1999). Suppression of ING1 expression in sporadic breast cancer. *Oncogene* **18**.
- 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>
- WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55** 2183–2202. [MR2729873](#) <https://doi.org/10.1109/TIT.2009.2016018>
- WANG, W., WAINWRIGHT, M. J. and RAMCHANDRAN, K. (2010). Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Trans. Inf. Theory* **56** 2967–2979. [MR2683451](#) <https://doi.org/10.1109/TIT.2010.2046199>
- YAN, J., JIANG, J.-Y., MENG, X.-N., XIU, Y.-L. and ZONG, Z.-H. (2016). MiR-23b targets cyclin G1 and suppresses ovarian cancer tumorigenesis and progression. *J. Exp. Clin. Cancer Res.* **35** 31.
- YANG, E., ALLEN, G., LIU, Z. and RAVIKUMAR, P. K. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems* 1358–1366.
- YANG, E., RAVIKUMAR, P. K., ALLEN, G. I. and LIU, Z. (2013). On Poisson graphical models. In *Advances in Neural Information Processing Systems* 1718–1726.
- YANG, E., RAVIKUMAR, P., ALLEN, G. I., BAKER, Y., WAN, Y.-W. and LIU, Z. (2014a). A general framework for mixed graphical models. ArXiv Preprint. Available at [arXiv:1411.0288](#).
- YANG, E., BAKER, Y., RAVIKUMAR, P., ALLEN, G. and LIU, Z. (2014b). Mixed graphical models via exponential families. In *Artificial Intelligence and Statistics* 1042–1050.
- YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16** 3813–3847. [MR3450553](#)
- YU, B. (2013). Stability. *Bernoulli* **19** 1484–1500. [MR3102560](#) <https://doi.org/10.3150/13-BEJSP14>
- YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. [MR2367824](#) <https://doi.org/10.1093/biomet/asm018>
- ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. [MR2604701](#) <https://doi.org/10.1214/09-AOS729>
- ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. [MR2435448](#) <https://doi.org/10.1214/07-AOS520>
- ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. [MR2274449](#)
- ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. [MR2279469](#) <https://doi.org/10.1198/016214506000000735>

MINING EVENTS WITH DECLASSIFIED DIPLOMATIC DOCUMENTS

BY YUANJUN GAO¹, JACK GOETZ², MATTHEW CONNELLY³ AND RAHUL MAZUMDER⁴

¹*Department of Statistics, Columbia University, gaoyuanjun0430@gmail.com*

²*Department of Statistics, University of Michigan, jackrgoetz@gmail.com*

³*MIT Sloan School of Management, Operations Research Center and Center for Statistics, MIT, rahulmaz@mit.edu*

⁴*Department of History, Columbia University, mjc96@columbia.edu*

Since 1973, the U.S. State Department has been using electronic record systems to preserve classified communications. Recently, approximately 1.9 million of these records from 1973–77 have been made available by the U.S. National Archives. While some of these communication streams have periods witnessing an acceleration in the rate of transmission, others do not show any notable patterns in communication intensity. Given the sheer volume of these communications, far greater than what had been available until now, scholars need automated statistical techniques to identify the communications that warrant closer study. We develop a statistical framework that can identify from a large corpus of documents a handful that historians would consider more interesting. Our approach brings together techniques from nonparametric signal estimation, statistical hypothesis testing and modern optimization methods—leading to a set of tools that help us identify and analyze various geometrical aspects of the communication streams. Dominant periods of heightened activities, as identified through these methods, correspond well with historical events recognized by standard reference works on the 1970s.

REFERENCES

- ATEFEH, F. and KHREICH, W. (2015). A survey of techniques for event detection in Twitter. *Comput. Intell.* **31** 132–164. [MR3318075 https://doi.org/10.1111/coin.12017](https://doi.org/10.1111/coin.12017)
- AUGER, I. E. and LAWRENCE, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.* **51** 39–54. [MR0978902 https://doi.org/10.1016/S0092-8240\(89\)80047-3](https://doi.org/10.1016/S0092-8240(89)80047-3)
- BECK, A. and TBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. [MR2486527 https://doi.org/10.1137/080716542](https://doi.org/10.1137/080716542)
- BEIELER, J., BRANDT, P. T., HALTERMAN, A., SCHRODT, P. A. and SIMPSON, E. M. (2016). Generating political event data in near real time: Opportunities and challenges. In *Computational Social Science* (R. M. Alvarez, ed.) 98–120. Cambridge Univ. Press, Cambridge. <https://doi.org/10.1017/CBO9781316257340.005>
- BELLMAN, R. and ROTH, R. (1969). Curve fitting by segmented straight lines. *J. Amer. Statist. Assoc.* **64** 1079–1084. [MR0246456](https://doi.org/10.2307/2284645)
- BERTSIMAS, D., KING, A. and MAZUMDER, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.* **44** 813–852. [MR3476618 https://doi.org/10.1214/15-AOS1388](https://doi.org/10.1214/15-AOS1388)
- BOYSEN, L., KEMPE, A., LIEBSCHER, V., MUNK, A. and WITTICH, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.* **37** 157–183. [MR2488348 https://doi.org/10.1214/07-AOS558](https://doi.org/10.1214/07-AOS558)
- BRODSKY, B. E. and DARKHOVSKY, B. S. (1993). *Nonparametric Methods in Change-Point Problems. Mathematics and Its Applications* **243**. Kluwer Academic, Dordrecht. [MR1228205 https://doi.org/10.1007/978-94-015-8163-9](https://doi.org/10.1007/978-94-015-8163-9)
- BRUCE, W., JENTLESON, PATERSON, T. G. and RIZOPOULOS, N. X. (1997). *Encyclopedia of US Foreign Relations* **2**. Oxford Univ. Press, New York.
- BRUNE, L. H. and BURNS, R. D. (2003). Chronological History of U.S. Foreign Relations: 1933–1988. Routledge. ISBN 9780415939164. Available at <https://books.google.com/books?id=nSe0gs-YaTkC>.
- DE CONDE, A., BURNS, R. D., LOGEVALL, F. and KETZ, L. B. (2002). *Encyclopedia of American Foreign Policy*. Scribner’s, New York.
- FLANDERS, S. A. and FLANDERS, C. N. (1993). Dictionary of American Foreign Affairs. Macmillan Library Reference.

- GAO, Y., GOETZ, J., CONNELLY, M. and MAZUMDER, R. (2020). Supplement to “Mining events with declassified diplomatic documents.” <https://doi.org/10.1214/20-AOAS1344SUPPA>, <https://doi.org/10.1214/20-AOAS1344SUPPB>
- GLAZ, J., NAUS, J. and WALLENSTEIN, S. (2001). *Scan Statistics. Springer Series in Statistics*. Springer, New York. MR1869112 <https://doi.org/10.1007/978-1-4757-3460-7>
- HANNA, A. (2014). Assessing gdel with handcoded protest data. (accessed: July 29, 2016). www.badhessian.org.
- HARMER, T. (2011). *Allende's Chile and the Inter-American Cold War*. Univ. of North Carolina Press, Chapel Hill, NC.
- 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>
- HAZIMEH, H. and MAZUMDER, R. (2018). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. Preprint. Available at [arXiv:1803.01454](https://arxiv.org/abs/1803.01454).
- HE, Q., CHANG, K. and LIM, E.-P. (2007). Analyzing feature trajectories for event detection. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 207–214. ACM, New York.
- JACKSON, B., SCARGLE, J. D., BARNES, D., ARABHI, S., ALT, A., GIOUMOUSIS, P., GWIN, E., SANG-TRAKULCHAROEN, P., TAN, L. et al. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Process. Lett.* **12** 105–108.
- JENKINS, J. C. and MAHER, T. V. (2016). What should we do about source selection in event data? Challenges, progress, and possible solutions. *Int. J. Sociol.* **46** 42–57.
- JOHNSON, N. A. (2013). A dynamic programming algorithm for the fused lasso and L_0 -segmentation. *J. Comput. Graph. Statist.* **22** 246–260. MR3173713 <https://doi.org/10.1080/10618600.2012.681238>
- KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418 <https://doi.org/10.1080/01621459.2012.737745>
- KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. MR2505584 <https://doi.org/10.1137/070690274>
- KLEINBERG, J. (2003). Bursty and hierarchical structure in streams. *Data Min. Knowl. Discov.* **7** 373–397. MR2011139 <https://doi.org/10.1023/A:1024940629314>
- LANGBART, D., FISCHER, W. and ROBERSON, L. (2007). The national weights and measures laboratory. appraisal of records covered by N1-59-07-3-P Technical report, Tech. rep, College Park: National Archives.
- MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. MR1429931 <https://doi.org/10.1214/aos/1034276635>
- MAZUMDER, R., FRIEDMAN, J. H. and HASTIE, T. (2011). *SparseNet*: Coordinate descent with nonconvex penalties. *J. Amer. Statist. Assoc.* **106** 1125–1138. MR2894769 <https://doi.org/10.1198/jasa.2011.tm09738>
- MAZUMDER, R., RADCHENKO, P. and DEDIEU, A. (2017). Subset selection with shrinkage: Sparse linear modeling when the snr is low. Preprint. Available at [arXiv:1708.03288](https://arxiv.org/abs/1708.03288).
- NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course. Applied Optimization* **87**. Kluwer Academic, Boston, MA. MR2142598 <https://doi.org/10.1007/978-1-4419-8853-9>
- NESTEROV, YU. (2013). Gradient methods for minimizing composite functions. *Math. Program.* **140** 125–161. MR3071865 <https://doi.org/10.1007/s10107-012-0629-5>
- OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics* **5** 557–572.
- PAGE, E. S. (1954). Continuous inspection schemes. *Biometrika* **41** 100–115. MR0088850 <https://doi.org/10.1093/biomet/41.1-2.100>
- PUI CHEONG FUNG, G., XU YU, J., YU, P. S. and LU, H. (2005). Parameter free bursty events detection in text streams. In *Proceedings of the 31st International Conference on Very Large Data Bases* 181–192. VLDB Endowment.
- SCOTT, A. J. and KNOTT, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* 507–512.
- TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323. MR3189487 <https://doi.org/10.1214/13-AOS1189>
- TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641 <https://doi.org/10.1111/j.1467-9868.2005.00490.x>
- TRUONG, C., OUDRE, L. and VAYATIS, N. (2018). A review of change point detection methods. CoRR. Available at [abs/1801.00718](https://arxiv.org/abs/1801.00718).
- WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689 <https://doi.org/10.1214/08-AOS646>

CLASSIFICATION FROM ONLY POSITIVE AND UNLABELED FUNCTIONAL DATA

BY YOSHIKAZU TERADA^{1,2}, ISSEI OGASAWARA^{3,*} AND KEN NAKATA^{3,†}

¹Graduate School of Engineering Science, Osaka University, terada@sigmath.es.osaka-u.ac.jp

²RIKEN Center for Advanced Intelligence Project (AIP)

³Graduate School of Medicine, Osaka University, *ogasawaraissei@hss.osaka-u.ac.jp; †ken-nakata@umin.ac.jp

In various fields, data recorded continuously during a time interval and curve data, such as spectral data, become common. These kinds of data can be interpreted as functional data. In this paper we have studied binary classification from only positive and unlabeled functional data (PU classification for functional data). Our first contribution is to present a simple classification algorithm for this problem. The key feature of the algorithm is that it is not required an estimation of the unknown class prior (or the constant probability that a positive object is labeled). It is worth noting that the idea of our method can be applied to kernel linear discriminant analysis for general data. Our second contribution is to prove that, under mild regularity conditions similar to those in a supervised context, the proposed algorithm can achieve perfect asymptotic classification in the context of PU classification. In fact, we show that the proposed algorithm works well not only in numerical experiments but also for real data examples. Moreover, as an important practical application, we have used the proposed algorithm to identify handball players at risk for anterior cruciate ligament (ACL) injury based on ground reaction force data.

REFERENCES

- BLANCHARD, G., LEE, G. and SCOTT, C. (2010). Semi-supervised novelty detection. *J. Mach. Learn. Res.* **11** 2973–3009. [MR2746544](#)
- BOSQ, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. *Lecture Notes in Statistics* **149**. Springer, New York. [MR1783138](#) <https://doi.org/10.1007/978-1-4612-1154-9>
- DELAIGLE, A. and HALL, P. (2012). Achieving near perfect classification for functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 267–286. [MR2899863](#) <https://doi.org/10.1111/j.1467-9868.2011.01003.x>
- DELAIGLE, A. and HALL, P. (2013). Classification using censored functional data. *J. Amer. Statist. Assoc.* **108** 1269–1283. [MR3174707](#) <https://doi.org/10.1080/01621459.2013.824893>
- DU PLESSIS, M. C., NIU, G. and SUGIYAMA, M. (2014). Analysis of learning from positive and unlabeled data. In *Proceedings of Advances in Neural Information Processing Systems* 27 703–711.
- DU PLESSIS, M. C., NIU, G. and SUGIYAMA, M. (2015). Convex formulation for learning from positive and unlabeled data. In *Proceedings of the 32nd International Conference on Machine Learning* 1386–1394.
- DU PLESSIS, M. C. and SUGIYAMA, M. (2014). Class prior estimation from positive and unlabeled data. *IEICE Trans. Inf. Syst.* **E97-D** 1358–1362.
- ELKAN, C. and NOTO, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 213–220.
- FEBRERO-BANDE, M., DE LA FUENTE, M. O., GALEANO, P., NIETO, A. and GARCÍA-PORTUGUÉS, E. (2014). *fda.usc: Functional data analysis and utilities for statistical computing*.
- FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis*. *Springer Series in Statistics*. Springer, New York. [MR2229687](#)
- HALL, P., MÜLLER, H.-G. and WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. [MR2278365](#) <https://doi.org/10.1214/009053606000000272>
- HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. *Springer Series in Statistics*. Springer, New York. [MR2920735](#) <https://doi.org/10.1007/978-1-4614-3655-3>

- HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley, Chichester. [MR3379106](#) <https://doi.org/10.1002/9781118762547>
- JAIN, S., WHITE, M., TROSSET, M. W. and RADIVOJAC, P. (2016). Nonparametric semi-supervised learning of class proportions. ArXiv.
- KALIVAS, J. H. (1997). Two data sets of near infrared spectra. *Chemom. Intell. Lab. Syst.* **37** 255–259.
- KAWANO, S. (2013). Semi-supervised logistic discrimination via labeled data and unlabeled data from different sampling distributions. *Stat. Anal. Data Min.* **6** 472–481. [MR3150893](#) <https://doi.org/10.1002/sam.11204>
- MENON, A. K., VAN ROOYEN, B., ONG, C. S. and WILLIAMSON, R. C. (2015). Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the 32nd International Conference on Machine Learning* 125–134.
- RAMSAY, J. O. and SILVERMAN, B. W. (2002). *Applied Functional Data Analysis*. Springer Series in Statistics. Springer, New York. [MR1910407](#) <https://doi.org/10.1007/b98886>
- RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer Series in Statistics. Springer, New York. [MR2168993](#)
- SCOTT, C. and BLANCHARD, G. (2009). Novelty detection: Unlabeled data definitely help. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics* 464–471.
- SHANG, H. L. and HYNDMAN, R. J. (2013). fds: Functional data sets.
- TERADA, Y., OGASAWARA, I. and NAKATA, K. (2020). Supplement to “Classification from only positive and unlabeled functional data.” <https://doi.org/10.1214/20-AOAS1404SUPP>
- WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.
- ZHANG, X. and WANG, J.-L. (2016). From sparse to dense functional data and beyond. *Ann. Statist.* **44** 2281–2321. [MR3546451](#) <https://doi.org/10.1214/16-AOS1446>

IDENTIFYING MAIN EFFECTS AND INTERACTIONS AMONG EXPOSURES USING GAUSSIAN PROCESSES

BY FEDERICO FERRARI^{*} AND DAVID B. DUNSON[†]

Department of Statistical Science, Duke University, ^{*}ff31@duke.edu; [†]dunson@duke.edu

This article is motivated by the problem of studying the joint effect of different chemical exposures on human health outcomes. This is essentially a nonparametric regression problem, with interest being focused not on a black box for prediction but instead on selection of main effects and interactions. For interpretability we decompose the expected health outcome into a linear main effect, pairwise interactions and a nonlinear deviation. Our interest is in model selection for these different components, accounting for uncertainty and addressing nonidentifiability between the linear and nonparametric components of the semiparametric model. We propose a Bayesian approach to inference, placing variable selection priors on the different components, and developing a Markov chain Monte Carlo (MCMC) algorithm. A key component of our approach is the incorporation of a heredity constraint to only include interactions in the presence of main effects, effectively reducing dimensionality of the model search. We adapt a projection approach developed in the spatial statistics literature to enforce identifiability in modeling the nonparametric component using a Gaussian process. We also employ a dimension reduction strategy to sample the nonlinear random effects that aids the mixing of the MCMC algorithm. The proposed MixSelect framework is evaluated using a simulation study, and is illustrated using data from the National Health and Nutrition Examination Survey (NHANES). Code is available on GitHub.

REFERENCES

- BANERJEE, A., DUNSON, D. B. and TOKDAR, S. T. (2013). Efficient Gaussian process regression for large datasets. *Biometrika* **100** 75–89. [MR3034325 https://doi.org/10.1093/biomet/ass068](https://doi.org/10.1093/biomet/ass068)
- 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](https://doi.org/10.1214/13-AOS1096)
- BOBB, J. F., VALERI, L., CLAUS HENN, B., CHRISTIANI, D. C., WRIGHT, R. O., MAZUMDAR, M., GODLESKI, J. J. and COULL, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* **16** 493–508. [MR3365442 https://doi.org/10.1093/biostatistics/kxu058](https://doi.org/10.1093/biostatistics/kxu058)
- BRAUN, J. M., GENNINGS, C. and HAUSER, R. (2016). What can epidemiological studies tell us about the impact of chemical mixtures on human health? *Environ. Health Perspect.* **124** A6–A9.
- BUMAN, M. P., WINKLER, E. A. H., KURKA, J. M., HEKLER, E. B., BALDWIN, C. M., OWEN, N., AINSWORTH, B. E., HEALY, G. N. and GARDINER, P. A. (2013). Reallocating time to sleep, sedentary behaviors, or active behaviors: Associations with cardiovascular disease risk biomarkers, NHANES 2005–2006. *Am. J. Epidemiol.* **179** 323–334. <https://doi.org/10.1093/aje/kwt292>
- CHIPMAN, H. (1996). Bayesian variable selection with related predictors. *Canad. J. Statist.* **24** 17–36. [MR1394738 https://doi.org/10.2307/3315687](https://doi.org/10.2307/3315687)
- CLAUS HENN, B., COULL, B. A. and WRIGHT, R. O. (2014). Chemical mixtures and children’s health. *Curr. Opin. Pediatr.* **26** 223–229.
- DOMINICI, F., PENG, ROGER, D., BARR, CHRISTOPHER, D. and BELL, M. L. (2010). Protecting human health from air pollution: Shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology* **21** 187.
- FERRARI, F. and DUNSON, D. B. (2020a). Supplement to “Identifying main effects and interactions among exposures using Gaussian processes.” <https://doi.org/10.1214/20-AOAS1363SUPPA>, <https://doi.org/10.1214/20-AOAS1363SUPPB>

Key words and phrases. Bayesian modeling, chemical mixtures, Gaussian process, interaction selection, semi-parametric, strong heredity, variable selection.

- FERRARI, F. and DUNSON, D. B. (2020b). Bayesian factor analysis for inference on interactions. *J. Amer. Statist. Assoc.* To appear. <https://doi.org/10.1080/01621459.2020.1745813>
- GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.
- GUAN, Y. and HARAN, M. (2018). A computationally efficient projection-based approach for spatial generalized linear mixed models. *J. Comput. Graph. Statist.* **27** 701–714. MR3890863 <https://doi.org/10.1080/10618600.2018.1425625>
- HANKS, E. M., SCHLIEP, E. M., HOOTEN, M. B. and HOETING, J. A. (2015). Restricted spatial regression in practice: Geostatistical models, confounding, and robustness under model misspecification. *Environmetrics* **26** 243–254. MR3340961 <https://doi.org/10.1002/env.2331>
- HAO, N., FENG, Y. and ZHANG, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *J. Amer. Statist. Assoc.* **113** 615–625. MR3832213 <https://doi.org/10.1080/01621459.2016.1264956>
- HAO, N. and ZHANG, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **109** 1285–1301. MR3265697 <https://doi.org/10.1080/01621459.2014.881741>
- HARIS, A., WITTEN, D. and SIMON, N. (2016). Convex modeling of interactions with strong heredity. *J. Comput. Graph. Statist.* **25** 981–1004. MR3572025 <https://doi.org/10.1080/10618600.2015.1067217>
- HARVILLE, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer, New York. MR1467237 <https://doi.org/10.1007/b98818>
- HU, W., Mengersen, K., McMichael, A. and Tong, S. (2008). Temperature, air pollution and total mortality during summers in Sydney, 1994–2004. *Int. J. Biometeorol.* **52** 689–696.
- KUNDU, S. and DUNSON, D. B. (2014). Latent factor models for density estimation. *Biometrika* **101** 641–654. MR3254906 <https://doi.org/10.1093/biomet/asu019>
- LAMPA, E., LIND, L., LIND, P. M. and BORNEFALK-HERMANSSON, A. (2014). The identification of complex interactions in epidemiology and toxicology: A simulation study of boosted regression trees. *Environ. Health* **13** 57. <https://doi.org/10.1186/1476-069X-13-57>
- LANG, S. and BREZGER, A. (2004). Bayesian P-splines. *J. Comput. Graph. Statist.* **13** 183–212. MR2044877 <https://doi.org/10.1198/1061860043010>
- LIU, S. H., BOBB, J. F., LEE, K. H. et al. (2018). Lagged kernel machine regression for identifying time windows of susceptibility to exposures of complex mixtures. *Biostatistics* **19** 325–341. MR3815175 <https://doi.org/10.1093/biostatistics/kxx036>
- LUNN, D., BEST, N., SPIEGELHALTER, D., GRAHAM, G. and NEUENSCHWANDER, B. (2009). Combining MCMC with ‘sequential’ PKPD modelling. *J. Pharmacokinet. Pharmacodyn.* **36** 19.
- LYNCH, B. M., DUNSTAN, D. W., HEALY, G. N., WINKLER, E., EAKIN, E. and OWEN, N. (2010). Objectively measured physical activity and sedentary time of breast cancer survivors, and associations with adiposity: Findings from NHANES (2003–2006). *Cancer Causes Control* **21** 283–288.
- MACLEHOSE, R. F., DUNSON, D. B., HERRING, A. H. and HOPPIN, J. A. (2007). Bayesian methods for highly correlated exposure data. *Epidemiology* **18** 199–207. <https://doi.org/10.1097/01.ede.0000256320.30737.c0>
- MAUDERLY, J. L. and SAMET, J. M. (2009). Is there evidence for synergy among air pollutants in causing health effects? *Environ. Health Perspect.* **117** 1–6. <https://doi.org/10.1289/ehp.11654>
- MAUDERLY, J. L., BURNETT, R. T., CASTILLEJOS, M., OZKAYNAK, H., SAMET, J. M., STIEB, D. M., VEDAL, S. and WYZGA, R. E. (2010). Is the air pollution health research community prepared to support a multipollutant air quality management framework? *Inhal. Toxicol.* **22** 1–19. <https://doi.org/10.3109/08958371003793846>
- MORTIMER, K., NEUGEBAUER, R., LURMANN, F. and ALCORN, S. (2008). Air pollution and pulmonary function in asthmatic children: Effects of prenatal and lifetime exposures. *Epidemiology* **19** 550–557.
- NAGELKERKE, N. J. D., BERNSEN, R. M. D., SGAIER, S. K. and JHA, P. (2006). Body mass index, sexual behaviour, and sexually transmitted infections: An analysis using the NHANES 1999–2000 data. *BMC Public Health* **6** 199.
- NATIONAL RESEARCH COUNCIL ET AL. (2004). *Research Priorities for Airborne Particulate Matter: IV. Continuing Research Progress* **4**. National Academies Press.
- PADILLA, M. A., ELOBEID, M., RUDEN, D. M. and ALLISON, D. B. (2010). An examination of the association of selected toxic metals with total and central obesity indices: NHANES 99–02. *Int. J. Environ. Res. Public Health* **7** 3332–3347.
- QI, Y. A., MINKA, T. P., PICARD, R. W. and GHAHRAMANI, Z. (2004). Predictive automatic relevance determination by expectation propagation. In *Proceedings of the Twenty-First International Conference on Machine Learning* 85 ACM, New York.
- REICH, B. J., FUENTES, M. and DUNSON, D. B. (2011). Bayesian spatial quantile regression. *J. Amer. Statist. Assoc.* **106** 6–20. MR2816698 <https://doi.org/10.1198/jasa.2010.ap09237>

- ROBERTS, S. and MARTIN, M. (2005). A critical assessment of shrinkage-based regression approaches for estimating the adverse health effects of multiple air pollutants. *Atmos. Environ.* **39** 6223–6230.
- SANDERS, A. P., CLAUS HENN, B. and WRIGHT, R. O. (2015). Perinatal and childhood exposure to cadmium, manganese, and metal mixtures and effects on cognition and behavior: A review of recent literature. *Curr. Environ. Health Rep.* **2** 284–294.
- SAVITSKY, T., VANNUCCI, M. and SHA, N. (2011). Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statist. Sci.* **26** 130–149. [MR2849913](#) <https://doi.org/10.1214/11-STS354>
- SHAO, W., LIU, Q., HE, X., LIU, H. and GU, A. (2017). Association between level of urinary trace heavy metals and obesity among children aged 6–19 years: NHANES 1999–2011. *Environ. Sci. Pollut. Res. Int.* **24** 11573–11581.
- SI, Y., PILLAI, N. S. and GELMAN, A. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Anal.* **10** 605–625. [MR3420817](#) <https://doi.org/10.1214/14-BA924>
- SINISI, S. E. and VAN DER LAAN, M. J. (2004). Deletion/substitution/addition algorithm in learning with applications in genomics. *Stat. Appl. Genet. Mol. Biol.* **3** Art. 18. [MR2101467](#) <https://doi.org/10.2202/1544-6115.1069>
- VALERI, L., MAZUMDAR, M. M., BOBB, J. F., CLAUS HENN, B., RODRIGUES, E., SHARIF, O. I. A., KILE, M. L., QUAMRUZZAMAN, Q., AFROZ, S. et al. (2017). The joint effect of prenatal exposure to metal mixtures on neurodevelopmental outcomes at 20–40 months of age: Evidence from rural Bangladesh. *Environ. Health Perspect.* **125** 067015. <https://doi.org/10.1289/EHP614>
- VEDAL, S. and KAUFMAN, J. D. (2011). What does multi-pollutant air pollution research mean? *Am. J. Respir. Crit. Care Med.* **183** 4–6. <https://doi.org/10.1164/rccm.201009-1520ED>
- WANG, C. and JIANG, B. (2019). Penalized interaction estimation for ultrahigh dimensional quadratic regression. Preprint. Available at [arXiv:1901.07147](https://arxiv.org/abs/1901.07147).
- YU, K. and MOYEED, R. A. (2001). Bayesian quantile regression. *Statist. Probab. Lett.* **54** 437–447. [MR1861390](#) [https://doi.org/10.1016/S0167-7152\(01\)00124-9](https://doi.org/10.1016/S0167-7152(01)00124-9)
- YU, K., CHEN, C. W. S., REED, C. and DUNSON, D. B. (2013). Bayesian variable selection in quantile regression. *Stat. Interface* **6** 261–274. [MR3066690](#) <https://doi.org/10.4310/SII.2013.v6.n2.a9>

EFFECTIVE MODEL CALIBRATION VIA SENSIBLE VARIABLE IDENTIFICATION AND ADJUSTMENT WITH APPLICATION TO COMPOSITE FUSELAGE SIMULATION

BY YAN WANG¹, XIAOWEI YUE², RUI TUO³, JEFFREY H. HUNT⁴ AND JIANJUN SHI⁵

¹College of Applied Sciences, Beijing University of Technology, yanwang@bjut.edu.cn

²Grado Department of Industrial & Systems Engineering, Virginia Polytechnic Institute and State University, xwy@vt.edu

³Wm Michael Barnes '64 Department of Industrial & Systems Engineering, Texas A&M University, ruituo@tamu.edu

⁴The Boeing company, jeffrey.h.hunt@boeing.com

⁵The H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, jianjun.shi@isye.gatech.edu

Estimation of model parameters of computer simulators, also known as calibration, is an important topic in many engineering applications. In this paper we consider the calibration of computer model parameters with the help of engineering design knowledge. We introduce the concept of sensible (calibration) variables. Sensible variables are model parameters, which are sensitive in the engineering modeling, and whose optimal values differ from the engineering design values. We propose an effective calibration method to identify and to determine appropriate levels for the sensible variables with limited physical experimental data. The methodology is applied to a composite fuselage simulation problem.

REFERENCES

- BA, S. and JOSEPH, V. R. (2012). Composite Gaussian process models for emulating expensive functions. *Ann. Appl. Stat.* **6** 1838–1860. [MR3058685 https://doi.org/10.1214/12-AOAS570](https://doi.org/10.1214/12-AOAS570)
- BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data. Monographs on Statistics and Applied Probability* **135**. CRC Press, Boca Raton, FL. [MR3362184](https://doi.org/10.1214/04-AM-135)
- BARBERO, E. J. (2013). *Finite Element Analysis of Composite Materials Using ANSYS®*. CRC Press, Boca Raton, FL.
- BAYARRI, M. J., BERGER, J. O., PAULO, R., SACKS, J., CAPEO, J. A., CAVENDISH, J., LIN, C.-H. and TU, J. (2007). A framework for validation of computer models. *Technometrics* **49** 138–154. [MR2380530 https://doi.org/10.1198/004017007000000092](https://doi.org/10.1198/004017007000000092)
- BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Series in Statistics*. Springer, Heidelberg. [MR2807761 https://doi.org/10.1007/978-3-642-20192-9](https://doi.org/10.1007/978-3-642-20192-9)
- CHANG, C.-J. and JOSEPH, V. R. (2014). Model calibration through minimal adjustments. *Technometrics* **56** 474–482. [MR3290617 https://doi.org/10.1080/00401706.2013.850113](https://doi.org/10.1080/00401706.2013.850113)
- CRAIG, P. S., GOLDSTEIN, M., ROUGIER, J. C. and SEHEULT, A. H. (2001). Bayesian forecasting for complex systems using computer simulators. *J. Amer. Statist. Assoc.* **96** 717–729. [MR1946437 https://doi.org/10.1198/016214501753168370](https://doi.org/10.1198/016214501753168370)
- FERABOLI, P., PEITSO, E., DELEO, F., CLEVELAND, T. and STICKLER, P. B. (2009). Characterization of prepreg-based discontinuous carbon fiber/epoxy systems. *J. Reinf. Plast. Compos.* **28** 1191–1214.
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Vol. 1. Springer Series in Statistics*. Springer, Berlin. [MR2722294 https://doi.org/10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7)
- GRAMACY, R. B., BINGHAM, D., HOLLOWAY, J. P., GROSSKOPF, M. J., KURANZ, C. C., RUTTER, E., TRAN-THAM, M. and DRAKE, P. R. (2015). Calibrating a large computer experiment simulating radiative shock hydrodynamics. *Ann. Appl. Stat.* **9** 1141–1168. [MR3418718 https://doi.org/10.1214/15-AOAS850](https://doi.org/10.1214/15-AOAS850)
- HIGDON, D., KENNEDY, M., CAVENDISH, J. C., CAPEO, J. A. and RYNE, R. D. (2004). Combining field data and computer simulations for calibration and prediction. *SIAM J. Sci. Comput.* **26** 448–466. [MR2116355 https://doi.org/10.1137/S1064827503426693](https://doi.org/10.1137/S1064827503426693)

- JONES, R. M. (1998). *Mechanics of Composite Materials*. CRC Press, Boca Raton, FL.
- JOSEPH, V. R. and YAN, H. (2015). Engineering-driven statistical adjustment and calibration. *Technometrics* **57** 257–267. [MR3369681](#) <https://doi.org/10.1080/00401706.2014.902773>
- KENNEDY, M. C. and O'HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. [MR1858398](#) <https://doi.org/10.1111/1467-9868.00294>
- MALLICK, P. K. (2007). *Fiber-Reinforced Composites: Materials, Manufacturing, and Design*. CRC Press, Boca Raton, FL.
- MORRIS, M. D. and MITCHELL, T. J. (1995). Exploratory designs for computational experiments. *J. Statist. Plann. Inference* **43** 381–402.
- 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., JOSEPH, V. R. and YANG, H. (2016). Calibrating functional parameters in the ion channel models of cardiac cells. *J. Amer. Statist. Assoc.* **111** 500–509. [MR3538682](#) <https://doi.org/10.1080/01621459.2015.1119695>
- POWELL, M. J. D. (2006). The NEWUOA software for unconstrained optimization without derivatives. In *Large-Scale Nonlinear Optimization. Nonconvex Optim. Appl.* **83** 255–297. Springer, New York. [MR2194578](#) https://doi.org/10.1007/0-387-30065-1_16
- PRATOLA, M. T. and HIGDON, D. M. (2016). Bayesian additive regression tree calibration of complex high-dimensional computer models. *Technometrics* **58** 166–179. [MR3488296](#) <https://doi.org/10.1080/00401706.2015.1049749>
- SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2013). *The Design and Analysis of Computer Experiments. Springer Series in Statistics*. Springer, New York. [MR3887662](#)
- SHI, J. (2006). *Stream of Variation Modeling and Analysis for Multistage Manufacturing Processes*. CRC Press, Boca Raton, FL.
- SOBOL, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simulation* **55** 271–280. [MR1823119](#) [https://doi.org/10.1016/S0378-4754\(00\)00270-6](https://doi.org/10.1016/S0378-4754(00)00270-6)
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- TUO, R. (2019). Adjustments to computer models via projected kernel calibration. *SIAM/ASA J. Uncertain. Quantificat.* **7** 553–578. [MR3947296](#) <https://doi.org/10.1137/17M1128769>
- 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>
- WANG, H., LI, G. and TSAI, C.-L. (2007a). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 63–78. [MR2301500](#) <https://doi.org/10.1111/j.1467-9868.2007.00577.x>
- WANG, H., LI, R. and TSAI, C.-L. (2007b). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. [MR2410008](#) <https://doi.org/10.1093/biomet/asm053>
- WANG, Y., YUE, X., TUO, R., HUNT, J. H. and SHI, J. (2020). Supplement to “Effective model calibration via sensible variable identification and adjustment with application to composite fuselage simulation.” <https://doi.org/10.1214/20-AOAS1353SUPP>
- YUE, X., WEN, Y., HUNT, J. H. and SHI, J. (2018). Surrogate model based control considering uncertainty for composite fuselage assembly. *ASME Trans. J. Manuf. Sci. Eng.* **140** 041017. <https://doi.org/10.1115/1.4038510>
- ZHANG, T. and SHI, J. (2016a). Stream of variation modeling and analysis for compliant composite part assembly—Part I: Single-station processes. *ASME Trans. J. Manuf. Sci. Eng.* **138** 121003.
- ZHANG, T. and SHI, J. (2016b). Stream of variation modeling and analysis for compliant composite part assembly—Part II: Multistation processes. *ASME Trans. J. Manuf. Sci. Eng.* **138** 121004.
- ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. [MR2279469](#) <https://doi.org/10.1198/016214506000000735>

A NEAREST-NEIGHBOR BASED NONPARAMETRIC TEST FOR VIRAL REMODELING IN HETEROGENEOUS SINGLE-CELL PROTEOMIC DATA

BY TRAMBAK BANERJEE¹, BHASWAR B. BHATTACHARYA² AND
GOURAB MUKHERJEE³

¹*Analytics, Information and Operations Management, School of Business, University of Kansas, trambak@ku.edu*

²*Department of Statistics, The Wharton School, University of Pennsylvania, bbh@wharton.upenn.edu*

³*Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, gmukherj@marshall.usc.edu*

An important problem in contemporary immunology studies based on single-cell protein expression data is to determine whether cellular expressions are remodeled postinfection by a pathogen. One natural approach for detecting such changes is to use nonparametric two-sample statistical tests. However, in single-cell studies direct application of these tests is often inadequate, because single-cell level expression data from processed uninfected populations often contain attributes of several latent subpopulations with highly heterogeneous characteristics. As a result, viruses often infect these different subpopulations at different rates, in which case the traditional nonparametric two-sample tests for checking similarity in distributions are no longer conservative. In this paper, we propose a new nonparametric method for *Testing Remodeling under Heterogeneity* (TRUH) that can accurately detect changes in the infected samples compared to possibly heterogeneous uninfected samples. Our testing framework is based on composite nulls and is designed to allow the null model to encompass the possibility that the infected samples, though unaltered by the virus, might be dominantly arising from underrepresented subpopulations in the baseline data. The TRUH statistic, which uses nearest neighbor projections of the infected samples into the baseline uninfected population, is calibrated using a novel bootstrap algorithm. We demonstrate the nonasymptotic performance of the test via simulation experiments and also derive the large sample limit of the test statistic which provides theoretical support toward consistent asymptotic calibration of the test. We use the TRUH statistic for studying remodeling in tonsillar T cells under different types of HIV infection and find that, unlike traditional tests which do not have any heterogeneity correction, TRUH based statistical inference conforms to the biologically validated immunological theories on HIV infection.

REFERENCES

- AMIR, E.-A. D., DAVIS, K. L., TADMOR, M. D., SIMONDS, E. F., LEVINE, J. H., BENDALL, S. C., SHENFELD, D. K., KRISHNASWAMY, S., NOLAN, G. P. et al. (2013). Visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31** 545–552.
- ASLAN, B. and ZECH, G. (2005). New test for the multivariate two-sample problem based on the concept of minimum energy. *J. Stat. Comput. Simul.* **75** 109–119. MR2117010 <https://doi.org/10.1080/00949650410001661440>
- BANERJEE, T., BHATTACHARYA, B. B. and MUKHERJEE, G. (2020). Supplement to “A nearest-neighbor based nonparametric test for viral remodeling in heterogeneous single-cell proteomic data.” <https://doi.org/10.1214/20-AOAS1362SUPPA>, <https://doi.org/10.1214/20-AOAS1362SUPPB>
- BARINGHAUS, L. and FRANZ, C. (2004). On a new multivariate two-sample test. *J. Multivariate Anal.* **88** 190–206. MR2021870 [https://doi.org/10.1016/S0047-259X\(03\)00079-4](https://doi.org/10.1016/S0047-259X(03)00079-4)
- BASMACIOGULLARI, S. and PIZZATO, M. (2014). The activity of nef on hiv-1 infectivity. *Front. Microbiol.* **5** 232.

Key words and phrases. Single-cell virology, immunology, two-sample tests, viral remodeling, homogeneous Poisson process, nearest neighbors, HIV infection, mass cytometry.

- BENDALL, S. C., SIMONDS, E. F., QIU, P., AMIR, E.-A. D., KRUTZIK, P. O., FINCK, R., BRUGGNER, R. V., MELAMED, R., TREJO, A. et al. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332** 687–696.
- BENDALL, S. C., DAVIS, K. L., AMIR, E.-A. D., TADMOR, M. D., SIMONDS, E. F., CHEN, T. J., SHENFELD, D. K., NOLAN, G. P. and PE’ER, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell* **157** 714–725.
- BERRETT, T. B. and SAMWORTH, R. J. (2019a). Efficient two-sample functional estimation and the super-oracle phenomenon. arXiv preprint. Available at [arXiv:1904.09347](https://arxiv.org/abs/1904.09347).
- BERRETT, T. B. and SAMWORTH, R. J. (2019b). Nonparametric independence testing via mutual information. *Biometrika* **106** 547–566. [MR3992389 https://doi.org/10.1093/biomet/asz024](https://doi.org/10.1093/biomet/asz024)
- BERRETT, T. B., SAMWORTH, R. J. and YUAN, M. (2019). Efficient multivariate entropy estimation via k -nearest neighbour distances. *Ann. Statist.* **47** 288–318. [MR3909934 https://doi.org/10.1214/18-AOS1688](https://doi.org/10.1214/18-AOS1688)
- BHATTACHARYA, B. B. (2019). A general asymptotic framework for distribution-free graph-based two-sample tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 575–602. [MR3961499](https://doi.org/10.1111/rssb.12499)
- BICKEL, P. J. (1968). A distribution free version of the Smirnov two sample test in the p -variate case. *Ann. Math. Stat.* **40** 1–23. [MR0256519 https://doi.org/10.1214/aoms/1177697800](https://doi.org/10.1214/aoms/1177697800)
- BRUGGNER, R. V., BODENMILLER, B., DILL, D. L., TIBSHIRANI, R. J. and NOLAN, G. P. (2014). Automated identification of stratifying signatures in cellular subpopulations. *Proc. Natl. Acad. Sci. USA* **111** E2770–E2777.
- CANNINGS, T. I., BERRETT, T. B. and SAMWORTH, R. J. (2020). Local nearest neighbour classification with applications to semi-supervised learning. *Ann. Statist.* **48** 1789–1814. [MR4124344 https://doi.org/10.1214/19-AOS1868](https://doi.org/10.1214/19-AOS1868)
- CAVROIS, M., BANERJEE, T., MUKHERJEE, G., RAMAN, N., HUSSEIN, R., RODRIGUEZ, B. A., VASQUEZ, J., SPITZER, M. H., LAZARUS, N. H. et al. (2017). Mass cytometric analysis of hiv entry, replication, and remodeling in tissue cd4+ t cells. *Cell Rep.* **20** 984–998.
- CHAUDHURI, R., LINDWASSER, O. W., SMITH, W. J., HURLEY, J. H. and BONIFACINO, J. S. (2007). Down-regulation of cd4 by human immunodeficiency virus type 1 nef is dependent on clathrin and involves direct interaction of nef with the ap2 clathrin adaptor. *J. Virol.* **81** 3877–3890.
- CHEN, H., CHEN, X. and SU, Y. (2018). A weighted edge-count two-sample test for multivariate and object data. *J. Amer. Statist. Assoc.* **113** 1146–1155. [MR3862346 https://doi.org/10.1080/01621459.2017.1307757](https://doi.org/10.1080/01621459.2017.1307757)
- CHEN, L., DOU, W. W. and QIAO, Z. (2013). Ensemble subsampling for imbalanced multivariate two-sample tests. *J. Amer. Statist. Assoc.* **108** 1308–1323. [MR3174710 https://doi.org/10.1080/01621459.2013.800763](https://doi.org/10.1080/01621459.2013.800763)
- CHEN, H. and FRIEDMAN, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *J. Amer. Statist. Assoc.* **112** 397–409. [MR3646580 https://doi.org/10.1080/01621459.2016.1147356](https://doi.org/10.1080/01621459.2016.1147356)
- CHUNG, J. H. and FRASER, D. A. (1958). Randomization tests for a multivariate two-sample problem. *J. Amer. Statist. Assoc.* **53** 729–735.
- COVER, T. and HART, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13** 21–27.
- CRESSIE, N. (1976). On the logarithms of high-order spacings. *Biometrika* **63** 343–355. [MR0428583 https://doi.org/10.1093/biomet/63.2.343](https://doi.org/10.1093/biomet/63.2.343)
- DEB, N. and SEN, B. (2019). Multivariate rank-based distribution-free nonparametric testing using measure transportation. arXiv preprint. Available at [arXiv:1909.08733](https://arxiv.org/abs/1909.08733).
- DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics (New York)* **31**. Springer, New York. [MR1383093 https://doi.org/10.1007/978-1-4612-0711-5](https://doi.org/10.1007/978-1-4612-0711-5)
- DVORKIN, D. (2012). Icmix: Layered and chained mixture models. R package version 0.3/r5.
- FRIEDMAN, J. H. and RAFSKY, L. C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7** 697–717. [MR0532236](https://doi.org/10.1214/aos/1177697800)
- GADAT, S., KLEIN, T. and MARTEAU, C. (2016). Classification in general finite dimensional spaces with the k -nearest neighbor rule. *Ann. Statist.* **44** 982–1009. [MR3485951 https://doi.org/10.1214/15-AOS1395](https://doi.org/10.1214/15-AOS1395)
- GARCIA, J. V. and MILLER, A. D. (1991). Serine phosphorylation-independent downregulation of cell-surface cd4 by nef. *Nature* **350** 508.
- GHOSAL, P. and SEN, B. (2019). Multivariate ranks and quantiles using optimal transportation and applications to goodness-of-fit testing. arXiv preprint. Available at [arXiv:1905.05340](https://arxiv.org/abs/1905.05340).
- GIESEN, C., WANG, H. A., SCHAPIRO, D., ZIVANOVIC, N., JACOBS, A., HATTENDORF, B., SCHÜFFLER, P. J., GROLIMUND, D., BUHMANN, J. M. et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11** 417.
- GORIA, M. N., LEONENKO, N. N., MERGEL, V. V. and NOVI INVERARDI, P. L. (2005). A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparametr. Stat.* **17** 277–297. [MR2129834 https://doi.org/10.1080/104852504200026815](https://doi.org/10.1080/104852504200026815)
- GRETTON, A., BORGMARDT, K. M., RASCH, M., SCHÖLKOPF, B. and SMOLA, A. J. (2007). A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems* 513–520.

- HALL, P. and TAJVIDI, N. (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika* **89** 359–374. [MR1913964](#) <https://doi.org/10.1093/biomet/89.2.359>
- HECKEL, R. and BÖLCSKEI, H. (2015). Robust subspace clustering via thresholding. *IEEE Trans. Inf. Theory* **61** 6320–6342. [MR3418967](#) <https://doi.org/10.1109/TIT.2015.2472520>
- HENZE, N. (1984). Über die Anzahl von Zufallspunkten mit typ-gleichem nächsten Nachbarn und einen multi-variaten Zwei-Stichproben-Test. *Metrika* **31** 259–273. [MR0773815](#) <https://doi.org/10.1007/BF01915210>
- HENZE, N. and PENROSE, M. D. (1999). On the multivariate runs test. *Ann. Statist.* **27** 290–298. [MR1701112](#) <https://doi.org/10.1214/aos/1018031112>
- HOLMES, S. and HUBER, W. (2018). *Modern Statistics for Modern Biology*. Cambridge Univ. Press, Cambridge.
- 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.
- HWANG, B., LEE, J. H. and BANG, D. (2018). Single-cell rna sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50** 1–14.
- JAITIN, D. A., KENIGSBERG, E., KEREN-SHAUL, H., ELEFANT, N., PAUL, F., ZARETSKY, I., MILDNER, A., COHEN, N., JUNG, S. et al. (2014). Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science* **343** 776–779.
- JIA, C., HU, Y., KELLY, D., KIM, J., LI, M. and ZHANG, N. R. (2017). Accounting for technical noise in differential expression analysis of single-cell rna sequencing data. *Nucleic Acids Res.* **45** 10978–10988.
- JIANG, H., SOHN, L. L., HUANG, H. and CHEN, L. (2018). Single cell clustering based on cell-pair differentiability correlation and variance analysis. *Bioinformatics* **34** 3684–3694.
- KOZACHENKO, L. F. and LEONENKO, N. N. (1987). A statistical estimate for the entropy of a random vector. *Problemy Peredachi Informatsii* **23** 9–16. [MR0908626](#)
- LINDERMAN, M. D., BJORNSON, Z., SIMONDS, E. F., QIU, P., BRUGGNER, R. V., SHEODE, K., MENG, T. H., PLEVITIS, S. K. and NOLAN, G. P. (2012). Cytospade: High-performance analysis and visualization of high-dimensional cytometry data. *Bioinformatics* **28** 2400–2401.
- MACK, Y. P. (1983). Rate of strong uniform convergence of k -NN density estimates. *J. Statist. Plann. Inference* **8** 185–192. [MR0720150](#) [https://doi.org/10.1016/0378-3758\(83\)90037-X](https://doi.org/10.1016/0378-3758(83)90037-X)
- MACK, Y. P. and ROSENBLATT, M. (1979). Multivariate k -nearest neighbor density estimates. *J. Multivariate Anal.* **9** 1–15. [MR0530638](#) [https://doi.org/10.1016/0047-259X\(79\)90065-4](https://doi.org/10.1016/0047-259X(79)90065-4)
- MATHESON, N. J., SUMNER, J., WALS, K., RAPITEANU, R., WEEKES, M. P., VIGAN, R., WEINELT, J., SCHINDLER, M., ANTROBUS, R. et al. (2015). Cell surface proteomic map of hiv infection reveals antagonism of amino acid metabolism by vpu and nef. *Cell Host Microbe* **18** 409–423.
- MICHEL, N., ALLESBACH, I., VENZKE, S., FACKLER, O. T. and KEPPLER, O. T. (2005). The nef protein of human immunodeficiency virus establishes superinfection immunity by a dual strategy to downregulate cell-surface ccr5 and cd4. *Curr. Biol.* **15** 714–723.
- PENROSE, M. D. and YUKICH, J. E. (2003). Weak laws of large numbers in geometric probability. *Ann. Appl. Probab.* **13** 277–303. [MR1952000](#) <https://doi.org/10.1214/aoap/1042765669>
- QIU, P. (2012). Inferring phenotypic properties from single-cell characteristics. *PLoS ONE* **7** e37038.
- ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 515–530. [MR2168202](#) <https://doi.org/10.1111/j.1467-9868.2005.00513.x>
- ROSS, T. M., ORAN, A. E. and CULLEN, B. R. (1999). Inhibition of hiv-1 progeny virion release by cell-surface cd4 is relieved by expression of the viral nef protein. *Curr. Biol.* **9** 613–621.
- SAMWORTH, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist.* **40** 2733–2763. [MR3097618](#) <https://doi.org/10.1214/12-AOS1049>
- SCHIFFMAN, C., LIN, C., SHI, F., CHEN, L., SOHN, L. and HUANG, H. (2017). Sideseq: A cell similarity measure defined by shared identified differentially expressed genes for single-cell rna sequencing data. *Stat. Biosci.* **9** 200–216.
- SCHILLING, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 799–806. [MR0860514](#)
- SEN, N., MUKHERJEE, G. and ARVIN, A. M. (2015). Single cell mass cytometry reveals remodeling of human t cell phenotypes by varicella zoster virus. *Methods* **90** 85–94.
- SEN, A., ROTHENBERG, M. E., MUKHERJEE, G., FENG, N., KALISKY, T., NAIR, N., JOHNSTONE, I. M., CLARKE, M. F. and GREENBERG, H. B. (2012). Innate immune response to homologous rotavirus infection in the small intestinal villous epithelium at single-cell resolution. *Proc. Natl. Acad. Sci. USA* **109** 20667–20672.
- SEN, N., MUKHERJEE, G., SEN, A., BENDALL, S. C., SUNG, P., NOLAN, G. P. and ARVIN, A. M. (2014). Single-cell mass cytometry analysis of human tonsil t cell remodeling by varicella zoster virus. *Cell Rep.* **8** 633–645.

- SHI, F. and HUANG, H. (2017). Identifying cell subpopulations and their genetic drivers from single-cell RNA-Seq data using a biclustering approach. *J. Comput. Biol.* **24** 663–674. [MR3671106](#) <https://doi.org/10.1089/cmb.2017.0049>
- SONG, P. X.-K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scand. J. Stat.* **27** 305–320. [MR1777506](#) <https://doi.org/10.1111/1467-9469.00191>
- SWIGUT, T., SHOHDY, N. and SKOWRONSKI, J. (2001). Mechanism for down-regulation of cd28 by nef. *EMBO J.* **20** 1593–1604.
- TIBSHIRANI, R. and WALTHER, G. (2005). Cluster validation by prediction strength. *J. Comput. Graph. Statist.* **14** 511–528. [MR2170199](#) <https://doi.org/10.1198/106186005X59243>
- VAN DER MAATEN, L. and HINTON, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** 2579–2605.
- VASICEK, O. (1976). A test for normality based on sample entropy. *J. Roy. Statist. Soc. Ser. B* **38** 54–59. [MR0420958](#)
- VASSENA, L., GIULIANI, E., KOPPENSTEINER, H., BOLDUAN, S., SCHINDLER, M. and DORIA, M. (2015). Hiv-1 nef and vpu interfere with l-selectin (cd62l) cell surface expression to inhibit adhesion and signaling in infected cd4+ t lymphocytes. *J. Virol.* **JVI-00611**.
- WANG, J., HUANG, M., TORRE, E., DUECK, H., SHAFFER, S., MURRAY, J., RAJ, A., LI, M. and ZHANG, N. R. (2018). Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. USA* **115** E6437–E6446. [MR3831463](#) <https://doi.org/10.1073/pnas.1721085115>
- WEISS, L. (1960). Two-sample tests for multivariate distributions. *Ann. Math. Stat.* **31** 159–164. [MR0119305](#) <https://doi.org/10.1214/aoms/1177705995>
- ZHANG, J. and CHEN, H. (2017). Graph-based two-sample tests for discrete data. arXiv preprint. Available at [arXiv:1711.04349](https://arxiv.org/abs/1711.04349).

THE STATISTICAL PERFORMANCE OF MATCHING-ADJUSTED INDIRECT COMPARISONS: ESTIMATING TREATMENT EFFECTS WITH AGGREGATE EXTERNAL CONTROL DATA

BY DAVID CHENG¹, RAJEEV AYYAGARI^{2,*} AND JAMES SIGNOROVITCH^{2,†}

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

²*Analysis Group, Inc., * rajeev.ayyagari@analysisgroup.com; † james.signorovitch@analysisgroup.com*

Indirect comparisons of treatment-specific outcomes across separate studies often inform decision making in the absence of head-to-head randomized comparisons. Differences in baseline characteristics between study populations may introduce confounding bias in such comparisons. Matching-adjusted indirect comparison (MAIC) (*Pharmacoeconomics* **28** (2010) 935–945) has been used to adjust for differences in observed baseline covariates when the individual patient-level data (IPD) are available for only one study and aggregate data (AGD) are available for the other study. The approach weights outcomes from the IPD using estimates of trial selection odds that balance baseline covariates between the IPD and AGD. With the increasing use of MAIC, there is a need for formal assessments of its statistical properties. In this paper we formulate identification assumptions for causal estimands that justify MAIC estimators. We then examine large sample properties and evaluate strategies for estimating standard errors without the full IPD from both studies. The finite-sample bias of MAIC and the performance of confidence intervals based on different standard error estimators are evaluated through simulations. The method is illustrated through an example comparing placebo arm and natural history outcomes in Duchenne muscular dystrophy.

REFERENCES

- ADJEI, A. A., CHRISTIAN, M. and IVY, P. (2009). Novel designs and end points for phase II clinical trials. *Clin. Cancer Res.* **15** 1866–1872.
- ALTMAN, D., SONG, F., SAKAROVITCH, C., DEEKS, J., D’AMICO, R., BRADBURN, M. and EASTWOOD, A. G. A. (2005). Indirect comparisons of competing interventions. *Health Technology Assessment*.
- BLAND, J. M. and ALTMAN, D. G. (2000). Statistics notes. The odds ratio. *BMJ* **320** 1468. <https://doi.org/10.1136/bmj.320.7247.1468>
- BROOKHART, M. A., SCHNEEWEISS, S., ROTHMAN, K. J., GLYNN, R. J., AVORN, J. and STÜRMER, T. (2006). Variable selection for propensity score models. *Am. J. Epidemiol.* **163** 1149–1156.
- BUCHER, H. C., GUYATT, G. H., GRIFFITH, L. E. and WALTER, S. D. (1997). The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J. Clin. Epidemiol.* **50** 683–691.
- CARO, J. J. and ISHAK, K. J. (2010). No head-to-head trial? Simulate the missing arms. *Pharmacoeconomics* **28** 957–967.
- CHENG, D., AYYAGARI, R. and SIGNOROVITCH, J. (2020). Supplement to “The statistical performance of matching-adjusted indirect comparisons: Estimating treatment effects with aggregate external control data.” <https://doi.org/10.1214/20-AOAS1359SUPP>
- COHEN, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, San Diego.
- DE LUNA, X., WAERNBAUM, I. and RICHARDSON, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika* **98** 861–875. MR2860329 <https://doi.org/10.1093/biomet/asr041>
- EMA CHMP (2006). Guideline on clinical trials in small populations. London: EMEA.
- EMA CHMP (2018). Assessment Report: Kymriah (International non-proprietary name: Tisagenlecleucel). Available at https://www.ema.europa.eu/en/documents/assessment-report/kymriah-epar-public-assessment-report_en.pdf. [Online; accessed 1-April-2020].

- EMERY, A. E., MUNTONI, F. and QUINLIVAN, R. C. (2015). *Duchenne Muscular Dystrophy*. OUP, Oxford.
- FAY, M. P. and GRAUBARD, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* **57** 1198–1206. MR1950428 <https://doi.org/10.1111/j.0006-341X.2001.01198.x>
- FDA (2001). Guidance for Industry. E 10 choice of control group and related issues in clinical trials. US Department of Health and Human Services, Federal Drug Administration.
- FDA (2006). Drug Approval Package: Myozyme (Alglucosidase Alfa). Available at https://www.accessdata.fda.gov/drugsatfda_docs/nda/2006/125141s000_MyozymeTOC.cfm. [Online; accessed 11-June-2019].
- FDA (2017). FDA Approves First Treatment for a Form of Batten Disease. Available at <https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm555613.htm>. [Online; accessed 11-June-2019].
- GOEMANS, N., VANDEN HAUWE, M., SIGNOROVITCH, J., SWALLOW, E., SONG, J. and PROJECT, C. T. A. (2016). Individualized prediction of changes in 6-minute walk distance for patients with Duchenne muscular dystrophy. *PLoS ONE* **11**.
- 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>
- HAJNUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20** 25–46.
- HARTMAN, E., GRIEVE, R., RAMSAHAI, R. and SEKHON, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *J. Roy. Statist. Soc. Ser. A* **178** 757–778. MR3348358 <https://doi.org/10.1111/rssa.12094>
- ICH (2000). Choice of control group and related issues in clinical trials E10.
- IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 243–263. MR3153941 <https://doi.org/10.1111/rssb.12027>
- ISHAK, K. J., PROSKOROVSKY, I. and BENEDICT, A. (2015). Simulation and matching-based approaches for indirect comparison of treatments. *Pharmacoeconomics* **33** 537–549.
- JANSEN, J. P., FLEURENCE, R., DEVINE, B., ITZLER, R., BARRETT, A., HAWKINS, N., LEE, K., BOERSMA, C., ANNEMANS, L. et al. (2011). Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: Report of the ISPOR task force on indirect treatment comparisons good research practices: Part 1. *Value Health* **14** 417–428.
- 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>
- KAUERMANN, G. and CARROLL, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *J. Amer. Statist. Assoc.* **96** 1387–1396. MR1946584 <https://doi.org/10.1198/016214501753382309>
- KONG, A. (1992). A note on importance sampling using standardized weights. Univ. Chicago, Dept. Statistics, Tech. Rep. 348.
- 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>
- LU, G. and ADES, A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Stat. Med.* **23** 3105–3124.
- 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>
- MAZZONE, E. S., CORATTI, G., SORMANI, M. P., MESSINA, S., PANE, M., D'AMICO, A., COLIA, G., FANELLI, L., BERARDINELLI, A. et al. (2016). Timed rise from floor as a predictor of disease progression in Duchenne muscular dystrophy: An observational study. *PLoS ONE* **11**.
- NEWAY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics, Vol. IV. Handbooks in Econom.* **2** 2111–2245. North-Holland, Amsterdam. MR1315971
- NIE, L., ZHANG, Z., RUBIN, D. and CHU, J. (2013). Likelihood reweighting methods to reduce potential bias in noninferiority trials which rely on historical data to make inference. *Ann. Appl. Stat.* **7** 1796–1813. MR3127969 <https://doi.org/10.1214/13-AOAS655>
- NIXON, R. M., BANSBACK, N. and BRENNAN, A. (2007). Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. *Stat. Med.* **26** 1237–1254. MR2345718 <https://doi.org/10.1002/sim.2624>
- PHILLIPPO, D., ADES, T., DIAS, S., PALMER, S., ABRAMS, K. and WELTON, N. (2016). NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submissions to NICE. (Technical Support Documents). Decision Support Unit, ScHARR, University of Sheffield: NICE Decision Support Unit.
- POCOCK, S. J. (1976). The combination of randomized and historical controls in clinical trials. *J. Chronic Dis.* **29** 175–188.

- PRENTICE, R. L. and PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–411. [MR0556730 https://doi.org/10.1093/biomet/66.3.403](https://doi.org/10.1093/biomet/66.3.403)
- RICOTTI, V., RIDOUT, D. A., PANE, M., MAIN, M., MAYHEW, A., MERCURI, E., MANZUR, A. Y. and MUNTONI, F. (2016). The NorthStar ambulatory assessment in Duchenne muscular dystrophy: Considerations for the design of clinical trials. *J Neurol Neurosurg Psychiatry* **87** 149–155.
- ROTNITZKY, A., LI, L. and LI, X. (2010). A note on overadjustment in inverse probability weighted estimation. *Biometrika* **97** 997–1001. [MR2746169 https://doi.org/10.1093/biomet/asq049](https://doi.org/10.1093/biomet/asq049)
- RUBIN, D. B. and THOMAS, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics* 249–264.
- SIGNOROVITCH, J. E., WU, E. Q., ANDREW, P. Y., GERRITS, C. M., KANTOR, E., BAO, Y., GUPTA, S. R. and MULANI, P. M. (2010). Comparative effectiveness without head-to-head trials. *Pharmacoeconomics* **28** 935–945.
- SIGNOROVITCH, J. E., SIKIRICA, V., ERDER, M. H., XIE, J., LU, M., HODGKINS, P. S., BETTS, K. A. and WU, E. Q. (2012). Matching-adjusted indirect comparisons: A new tool for timely comparative effectiveness research. *Value Health* **15** 940–947.
- SIGNOROVITCH, J., SWALLOW, E., KANTOR, E., WANG, X., KLIMOVSKY, J., HAAS, T., DEVINE, B. and METRAKOS, P. (2013). Everolimus and sunitinib for advanced pancreatic neuroendocrine tumors: A matching-adjusted indirect comparison. *Experimental Hematology & Oncology* **2** 32.
- SNAPINN, S. and JIANG, Q. (2011). Indirect comparisons in the comparative efficacy and non-inferiority settings. *Pharm. Stat.* **10** 420–426.
- 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)
- STUART, E. A., COLE, S. R., BRADSHAW, C. P. and LEAF, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *J. Roy. Statist. Soc. Ser. A* **174** 369–386. [MR2898850 https://doi.org/10.1111/j.1467-985X.2010.00673.x](https://doi.org/10.1111/j.1467-985X.2010.00673.x)
- SUTTON, A., ADES, A., COOPER, N. and ABRAMS, K. (2008). Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics* **26** 753–767.
- SWALLOW, E., SONG, J., YUAN, Y., KALSEKAR, A., KELLEY, C., PEEPLES, M., MU, F., ACKERMAN, P. and SIGNOROVITCH, J. (2016). Daclatasvir and sofosbuvir versus sofosbuvir and ribavirin in patients with chronic hepatitis C coinfecting with HIV: A matching-adjusted indirect comparison. *Clinical Therapeutics* **38** 404–412.
- 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)
- WELLS, G., SULTAN, S., CHEN, L., KHAN, M. and COYLE, D. (2009). Indirect evidence: Indirect treatment comparisons in meta-analysis. Ottawa: Canadian Agency for Drugs and Technologies in Health 1–94.
- ZEILEIS, A. (2004). Econometric computing with HC and HAC covariance matrix estimators.
- ZHANG, Z., NIE, L., SOON, G. and HU, Z. (2016). New methods for treatment effect calibration, with applications to non-inferiority trials. *Biometrics* **72** 20–29. [MR3500570 https://doi.org/10.1111/biom.12388](https://doi.org/10.1111/biom.12388)
- ZHAO, Q. and PERCIVAL, D. (2017). Entropy balancing is doubly robust. *J. Causal Inference* **5**.
- ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* **110** 910–922. [MR3420672 https://doi.org/10.1080/01621459.2015.1023805](https://doi.org/10.1080/01621459.2015.1023805)

MIXTURE OF HIDDEN MARKOV MODELS FOR ACCELEROMETER DATA

BY MARIE DU ROY DE CHAUMARAY^{*}, MATTHIEU MARBAC[†] AND FABIEN NAVARRO[‡]

CREST, ENSAI, ^{*}marie.du-roy-de-chaumaray@ensai.fr; [†]matthieu.marbac-lourdelle@ensai.fr; [‡]fabien.navarro@ensai.fr

Motivated by the analysis of accelerometer data taken across a population of individuals, we introduce a specific finite mixture of hidden Markov models with particular characteristics that adapt well to the specific nature of this type of longitudinal data. Our model allows for the computation of statistics that characterize the physical activity of a subject (e.g., the mean time spent at different activity levels and the probability of the transition between two activity levels) without specifying the activity levels in advance but by estimating them from the data. In addition, this approach allows the heterogeneity of the population to be taken into account and subpopulations with homogeneous physical activity behavior to be defined. We prove that, under mild assumptions, this model implies that the probability of misclassifying a subject decreases at an exponential decay with the length of its measurement sequence. Model identifiability is also investigated. We also report a comprehensive suite of numerical simulations to support our theoretical findings. The method is motivated by and applied to the Physical Activity and Transit Survey.

REFERENCES

- AARTS, E. (2019). mHMMbayes: Multilevel hidden Markov models using Bayesian estimation. R package version 0.1.1.
- ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. [MR2549554](#) <https://doi.org/10.1214/09-AOS689>
- ALTMAN, R. M. (2007). Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting. *J. Amer. Statist. Assoc.* **102** 201–210. [MR2345538](#) <https://doi.org/10.1198/016214506000001086>
- BAI, J., SUN, Y., SCHRACK, J. A., CRAINICEANU, C. M. and WANG, M.-C. (2018). A two-stage model for wearable device data. *Biometrics* **74** 744–752. [MR3825361](#) <https://doi.org/10.1111/biom.12781>
- BANFIELD, J. D. and RAFTERY, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803–821. [MR1243494](#) <https://doi.org/10.2307/2532201>
- BARTOLUCCI, F., FARCOMENI, A. and PENNONI, F. (2013). *Latent Markov Models for Longitudinal Data. Statistics in the Social and Behavioral Sciences Series*. CRC Press, Boca Raton, FL. [MR3184304](#)
- BARTOLUCCI, F., PANDOLFI, S. and PENNONI, F. (2017). Lmest: An R package for latent Markov models for longitudinal categorical data. *J. Stat. Softw.* **81** 1–38.
- BARTOLUCCI, F., PENNONI, F. and VITTADINI, G. (2011). Assessment of school performance through a multi-level latent Markov rasch model. *J. Educ. Behav. Stat.* **36** 491–522.
- BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 719–725.
- BRAULT, V. and MARIADASSOU, M. (2015). Co-clustering through latent bloc model: A review. *J. SFdS* **156** 120–139. [MR3432606](#)
- CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in Hidden Markov Models. Springer Series in Statistics*. Springer, New York. With Randal Douc’s contributions to Chapter 9 and Christian P. Robert’s to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat. [MR2159833](#)
- CELEUX, G. and DURAND, J.-B. (2008). Selecting hidden Markov model state number with cross-validated likelihood. *Comput. Statist.* **23** 541–564. [MR2448181](#) <https://doi.org/10.1007/s00180-007-0097-1>

- CELISSE, A., DAUDIN, J.-J. and PIERRE, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.* **6** 1847–1899. MR2988467 <https://doi.org/10.1214/12-EJS729>
- COLE, R. J., KRIPKE, D. F., GRUEN, W., MULLANEY, D. J. and GILLIN, J. C. (1992). Automatic sleep/wake identification from wrist activity. *Sleep* **15** 461–469.
- CSISZÁR, I. and TALATA, Z. (2006). Consistent estimation of the basic neighborhood of Markov random fields. *Ann. Statist.* **34** 123–145. MR2275237 <https://doi.org/10.1214/009053605000000912>
- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537
- DU ROY DE CHAUMARAY, M., MARBAC, M. and NAVARRO, F. (2020b). Supplement to “Mixture of hidden Markov models for accelerometer data.” <https://doi.org/10.1214/20-AOAS1375SUPP>
- DU ROY DE CHAUMARAY, M., MARBAC, M. and NAVARRO, F. (2020a). *MHMM: Finite mixture of hidden Markov model*. R package version 1.0.0.
- DYRSTAD, S. M., HANSEN, B. H., HOLME, I. M. and ANDERSSSEN, S. A. (2014). Comparison of self-reported versus accelerometer-measured physical activity. *Med. Sci. Sports Exerc.* **46** 99–106. <https://doi.org/10.1249/MSS.0b013e3182a0595f>
- FREEDSON, P. S., MELANSON, E. and SIRARD, J. (1998). Calibration of the computer science and applications. *Inc. Accelerometer. Med. Sci. Sports Exerc.* **30** 777–781.
- GASSIAT, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. Henri Poincaré (B)* **38** 897–906. MR1955343 [https://doi.org/10.1016/S0246-0203\(02\)01125-1](https://doi.org/10.1016/S0246-0203(02)01125-1)
- GASSIAT, E., CLEYNEN, A. and ROBIN, S. (2016). Inference in finite state space non parametric hidden Markov models and applications. *Stat. Comput.* **26** 61–71. MR3439359 <https://doi.org/10.1007/s11222-014-9523-8>
- GERACI, M. (2019). Additive quantile regression for clustered data with an application to children’s physical activity. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 1071–1089. MR4002384
- GERACI, M. and FARCOMENI, A. (2016). Probabilistic principal component analysis to identify profiles of physical activity behaviours in the presence of non-ignorable missing data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 51–75. MR3438238 <https://doi.org/10.1111/rssc.12105>
- GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231. MR0370936 <https://doi.org/10.1093/biomet/61.2.215>
- GRANDNER, M. A., SANDS-LINCOLN, M. R., PAK, V. M. and GARLAND, S. N. (2013). Sleep duration, cardiovascular disease, and proinflammatory biomarkers. *Nat. Sci. Sleep* **5** 93–107. <https://doi.org/10.2147/NSS.S31063>
- GRUEN, M. E., ALFARO-CÓRDOBA, M., THOMSON, A. E., WORTH, A. C., STAICU, A.-M. and LASCELLES, B. D. X. (2017). The use of functional data analysis to evaluate activity in a spontaneous model of degenerative joint disease associated pain in cats. *PLoS ONE* **12** e0169576.
- HELSKE, S. and HELSKE, J. (2019). Mixture hidden Markov models for sequence data: The seqHMM package in R. *J. Stat. Softw.* **88** 1–32.
- HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262 <https://doi.org/10.1198/016214502388618906>
- HUANG, Q., COHEN, D., KOMARZYNSKI, S., LI, X.-M., INNOMINATO, P., LÉVI, F. and FINKENSTÄDT, B. (2018). Hidden Markov models for monitoring circadian rhythmicity in telemetric activity data. *J. R. Soc. Interface* **15** 20170885.
- HUANG, L., BAI, J., IVANESCU, A., HARRIS, T., MAURER, M., GREEN, P. and ZIPUNNIKOV, V. (2019). Multilevel matrix-variate analysis and its application to accelerometry-measured physical activity in clinical populations. *J. Amer. Statist. Assoc.* **114** 553–564. MR3963162 <https://doi.org/10.1080/01621459.2018.1482750>
- HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.
- HUNT, L. and JORGENSEN, M. (2011). Clustering mixed data. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **1** 352–361.
- HUNTER, D. R., GOODREAU, S. M. and HANDCOCK, M. S. (2008). Goodness of fit of social network models. *J. Amer. Statist. Assoc.* **103** 248–258. MR2394635 <https://doi.org/10.1198/016214507000000446>
- IMMERWAHR, S., WYKER, B., BARTLEY, K. and EISENHOWER, D. (2012). The physical activity and transit survey device follow-up study: Methodology report. The New York City Department of Health and Mental Hygiene.
- INNERD, P., HARRISON, R. and COULSON, M. (2018). Using open source accelerometer analysis to assess physical activity and sedentary behaviour in overweight and obese adults. *BMC Public Health* **18** 543. <https://doi.org/10.1186/s12889-018-5215-1>
- KARLIS, D. and MELIGKOTSIDOU, L. (2007). Finite mixtures of multivariate Poisson distributions with application. *J. Statist. Plann. Inference* **137** 1942–1960. MR2323875 <https://doi.org/10.1016/j.jspi.2006.07.001>

- KIMM, S. Y., GLYNN, N. W., OBARZANEK, E., KRISKA, A. M., DANIELS, S. R., BARTON, B. A. and LIU, K. (2005). Relation between the changes in physical activity and body-mass index during adolescence: A multi-centre longitudinal study. *Lancet* **366** 301–307.
- KOSMIDIS, I. and KARLIS, D. (2016). Model-based clustering using copulas with applications. *Stat. Comput.* **26** 1079–1099. [MR3535046 https://doi.org/10.1007/s11222-015-9590-5](https://doi.org/10.1007/s11222-015-9590-5)
- LEE, J. A. and GILL, J. (2018). Missing value imputation for physical activity data measured by accelerometer. *Stat. Methods Med. Res.* **27** 490–506. [MR3761118 https://doi.org/10.1177/0962280216633248](https://doi.org/10.1177/0962280216633248)
- LEE, I.-M., SHIROMA, E. J., LOBELO, F., PUSKA, P., BLAIR, S. N., KATZMARZYK, P. T. et al. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy. *Lancet* **380** 219–229.
- LEVIN, D. A. and PERES, Y. (2017). *Markov Chains and Mixing Times*. Amer. Math. Soc., Providence, RI. [MR3726904](https://doi.org/10.1090/S0025-5718-2017-076904-0)
- LIM, Y., OH, H.-S. and CHEUNG, Y. K. (2019). Functional clustering of accelerometer data via transformed input variables. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 495–520. [MR3937460](https://doi.org/10.1111/rssc.12460)
- LIM, S., WYKER, B., BARTLEY, K. and EISENHOWER, D. (2015). Measurement error of self-reported physical activity levels in New York city: Assessment and correction. *Am. J. Epidemiol.* **181** 648–655.
- MARUOTTI, A. (2011). Mixed hidden Markov models for longitudinal data: An overview. *Int. Stat. Rev.* **79** 427–454.
- MATIAS, C., REBAFKA, T. and VILLERS, F. (2018). A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika* **105** 665–680. [MR3842891 https://doi.org/10.1093/biomet/asy016](https://doi.org/10.1093/biomet/asy016)
- MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley, New York. [MR1789474 https://doi.org/10.1002/0471721182](https://doi.org/10.1002/0471721182)
- MCNICHOLAS, P. D. (2017). *Mixture Model-Based Classification*. CRC Press, Boca Raton, FL. [MR3642443](https://doi.org/10.1002/9781119464443)
- MCTIERNAN, A. (2008). Mechanisms linking physical activity with cancer. *Nat. Rev. Cancer* **8** 205–211. <https://doi.org/10.1038/nrc2325>
- MORRIS, J. S., ARROYO, C., COULL, B. A., RYAN, L. M., HERRICK, R. and GORTMAKER, S. L. (2006). Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: A case study. *J. Amer. Statist. Assoc.* **101** 1352–1364. [MR2307570 https://doi.org/10.1198/016214506000000465](https://doi.org/10.1198/016214506000000465)
- NOEL, S. E., MATTOCKS, C., EMMETT, P., RIDDOCH, C. J., NESS, A. R. and NEWBY, P. (2010). Use of accelerometer data in prediction equations for capturing implausible dietary intakes in adolescents. *Am. J. Clin. Nutr.* **92** 1436–1445.
- PALTA, P., MCMURRAY, R., GOUSKOVA, N., SOTRES-ALVAREZ, D., DAVIS, S., CARNETHON, M., CASTAÑEDA, S., GELLMAN, M., HANKINSON, A. L. et al. (2015). Self-reported and accelerometer-measured physical activity by body mass index in us Hispanic/latino adults: Hchs/sol. *Preventive medicine reports* **2** 824–828.
- POLLAK, C. P., TRYON, W. W., NAGARAJA, H. and DZWONCZYK, R. (2001). How accurately does wrist actigraphy identify the states of sleep and wakefulness? *Sleep* **24** 957–965.
- SADEH, A., SHARKEY, M. and CARSKADON, M. A. (1994). Activity-based sleep-wake identification: An empirical test of methodological issues. *Sleep* **17** 201–207.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. [MR0468014](https://doi.org/10.1214/aoms/1177703862)
- SCOTT, S. L., JAMES, G. M. and SUGAR, C. A. (2005). Hidden Markov models for longitudinal comparisons. *J. Amer. Statist. Assoc.* **100** 359–369. [MR2170459 https://doi.org/10.1198/016214504000001592](https://doi.org/10.1198/016214504000001592)
- SLOOTMAKER, S. M., SCHUIT, A. J., CHINAPAW, M. J., SEIDELL, J. C. and VAN MECHELEN, W. (2009). Disagreement in physical activity assessed by accelerometer and self-report in subgroups of age, gender, education and weight status. *Int. J. Behav. Nutr. Phys. Act.* **6** 17.
- TAHERI, S., LIN, L., AUSTIN, D., YOUNG, T. and MIGNOT, E. (2004). Short sleep duration is associated with reduced leptin, elevated ghrelin, and increased body mass index. *PLoS Med.* **1** e62.
- TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Stat.* **34** 1265–1269. [MR0155376 https://doi.org/10.1214/aoms/1177703862](https://doi.org/10.1214/aoms/1177703862)
- TEICHER, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Stat.* **38** 1300–1302. [MR0216635 https://doi.org/10.1214/aoms/1177698805](https://doi.org/10.1214/aoms/1177698805)
- TITSIAS, M. K., HOLMES, C. C. and YAU, C. (2016). Statistical inference in hidden Markov models using *k*-segment constraints. *J. Amer. Statist. Assoc.* **111** 200–215. [MR3494653 https://doi.org/10.1080/01621459.2014.998762](https://doi.org/10.1080/01621459.2014.998762)
- TROIANO, R. P., BERRIGAN, D., DODD, K. W., MASSE, L. C., TILERT, T. and MCDOWELL, M. (2008). Physical activity in the United States measured by accelerometer. *Med. Sci. Sports Exerc.* **40** 181–188.
- US DEPARTMENT OF HEALTH AND HUMAN SERVICES (2008). 2008 physical activity guidelines for americans: Be active, healthy, and happy! Available at <http://www.health.gov/paguidelines>.

- VAN HEES, V. T., SABIA, S., ANDERSON, K. N., DENTON, S. J., OLIVER, J., CATT, M., ABELL, J. G., KIVIMÄKI, M., TRENELL, M. I. et al. (2015). A novel, open access method to assess sleep duration using a wrist-worn accelerometer. *PLoS ONE* **10** e0142533. <https://doi.org/10.1371/journal.pone.0142533>
- VAN DE POL, F. and LANGEHEINE, R. (1990). Mixed Markov latent class models. *Sociol. Method.* 213–247.
- VITERBI, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *J. Amer. Statist. Assoc.* **13** 260–269.
- WALLACE, M. L., BUYSSE, D. J., GERMAIN, A., HALL, M. H. and IYENGAR, S. (2018). Variable selection for skewed model-based clustering: Application to the identification of novel sleep phenotypes. *J. Amer. Statist. Assoc.* **113** 95–110. **MR3803442** <https://doi.org/10.1080/01621459.2017.1330202>
- WITOWSKI, V., FORAITA, R., PITSILADIS, Y., PIGEOT, I. and WIRSIK, N. (2014). Using hidden Markov models to improve quantifying physical activity in accelerometer data—a simulation study. *PLoS ONE* **9** e114089. <https://doi.org/10.1371/journal.pone.0114089>
- WONG, C. S. and LI, W. K. (2000). On a mixture autoregressive model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 95–115. **MR1747398** <https://doi.org/10.1111/1467-9868.00222>
- WYKER, B., BARTLEY, K., HOLDER-HAYES, E., IMMERWAHR, S., EISENHOWER, D. and HARRIS, T. (2013). Self-reported and accelerometer-measured physical activity: a comparison in new york city. New York (NY): New York City Department of Health and Mental Hygiene: Epi Research Report, 1–12.
- XIAO, L., HUANG, L., SCHRACK, J. A., FERRUCCI, L., ZIPUNNIKOV, V. and CRAINICEANU, C. M. (2015). Quantifying the lifetime circadian rhythm of physical activity: A covariate-dependent functional approach. *Biostatistics* **16** 352–367. **MR3365433** <https://doi.org/10.1093/biostatistics/kxu045>
- YANG, C.-C. and HSU, Y.-L. (2010). A review of accelerometry-based wearable motion detectors for physical activity monitoring. *Sensors* **10** 7772–7788.

RNDCLONE: TUMOR SUBCLONE RECONSTRUCTION BASED ON INTEGRATING DNA AND RNA SEQUENCE DATA

BY TIANJIAN ZHOU^{1,*}, SUBHAJIT SENGUPTA², PETER MÜLLER³ AND YUAN JI^{1,†}

¹Department of Public Health Sciences, University of Chicago, *tjzhou95@gmail.com; †yji@health.bsd.uchicago.edu

²Research Institute, NorthShore University HealthSystem, subhajit06@gmail.com

³Department of Mathematics, University of Texas at Austin, pmueller@math.utexas.edu

Tumor cell population consists of genetically heterogeneous subpopulations, known as subclones. Bulk sequencing data using high-throughput sequencing technology provide total and variant DNA and RNA read counts for many nucleotide loci as a mixture of signals from different subclones. We present RNDClone as a tool to deconvolute the mixture and reconstruct the subclones with distinct DNA genotypes and RNA expression profiles. In particular, we infer the number and population frequencies of subclones as well as subclonal copy numbers, variant allele numbers and gene expression levels by jointly modeling DNA and RNA read counts from the same tumor samples based on generalized latent factor models. Incorporating data at the RNA level provides new insights into intra-tumor heterogeneity in addition to the existing DNA-based inference. Performance of RNDClone is assessed using simulated and real-world datasets, including an analysis of three samples from a lung cancer patient in The Cancer Genome Atlas (TCGA). A potential fatal subclone is identified from the primary tumor which could explain the rapid prognosis and sudden death of the patient despite a promising diagnosis by conventional standards. The R package RNDClone is available in the Supplementary Material (Zhou et al. (2020)) and online at <https://github.com/tianjianzhou/RNDClone>.

REFERENCES

- BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. [MR2806429 https://doi.org/10.1093/biomet/asr013](https://doi.org/10.1093/biomet/asr013)
- CAMPBELL, K. R., STEIF, A., LAKS, E., ZAHN, H., LAI, D., MCPHERSON, A., FARAHANI, H., KABEER, F., O’FLANAGAN, C. et al. (2019). Clonealign: Statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biol.* **20** 54.
- CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. [MR2655722 https://doi.org/10.1198/016214508000000869](https://doi.org/10.1198/016214508000000869)
- CIBULSKIS, K., LAWRENCE, M. S., CARTER, S. L., SIVACHENKO, A., JAFFE, D., SOUGNEZ, C., GABRIEL, S., MEYERSON, M., LANDER, E. S. et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31** 213–219.
- DESHWAR, A. G., VEMBU, S., YUNG, C. K., JANG, G. H., STEIN, L. and MORRIS, Q. (2015). PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16** 35. <https://doi.org/10.1186/s13059-015-0602-8>
- FAN, X., ZHOU, W., CHONG, Z., NAKHLEH, L. and CHEN, K. (2014). Towards accurate characterization of clonal heterogeneity based on structural variation. *BMC Bioinform.* **15** 299.
- GAO, C., BROWN, C. D. and ENGELHARDT, B. E. (2013). A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects. Preprint. Available at [arXiv:1310.4792](https://arxiv.org/abs/1310.4792).
- GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
- GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. [MR3235677](https://doi.org/10.1201/b16006)

Key words and phrases. Copy number, gene expression, high-throughput sequencing, intra-tumor heterogeneity, latent factor model, somatic mutation.

- GEYER, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface* 156–163. Interface Foundation of North America, Fairfax Station, VA.
- GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. [MR1380810 https://doi.org/10.1093/biomet/82.4.711](https://doi.org/10.1093/biomet/82.4.711)
- GRIFFITHS, T. L. and GHARAMANI, Z. (2011). The Indian buffet process: An introduction and review. *J. Mach. Learn. Res.* **12** 1185–1224. [MR2804598](https://doi.org/10.1214/ss/1009212673)
- HEPPNER, G. H. (1984). Tumor heterogeneity. *Cancer Res.* **44** 2259–2265.
- 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)
- ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. [MR1952729 https://doi.org/10.1198/016214501750332758](https://doi.org/10.1198/016214501750332758)
- JAIN, S. and NEAL, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *J. Comput. Graph. Statist.* **13** 158–182. [MR2044876 https://doi.org/10.1198/1061860043001](https://doi.org/10.1198/1061860043001)
- KLAMBAUER, G., SCHWARZBAUER, K., MAYR, A., CLEVERT, D.-A., MITTERECKER, A., BODENHOFER, U. and HOCHREITER, S. (2012). cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* **40** e69. <https://doi.org/10.1093/nar/gks003>
- KUIPERS, J., JAHN, K. and BEERENWINKEL, N. (2017). Advances in understanding tumour evolution through single-cell sequencing. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **1867** 127–138.
- LANDAU, D. A., CARTER, S. L., STOJANOV, P., MCKENNA, A., STEVENSON, K., LAWRENCE, M. S., SOUGNEZ, C., STEWART, C., SIVACHENKO, A. et al. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152** 714–726.
- LEE, J., MÜLLER, P., SENGUPTA, S., GULUKOTA, K. and JI, Y. (2016). Bayesian inference for intratumour heterogeneity in mutations and copy number variation. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 547–563. [MR3522952 https://doi.org/10.1111/rssc.12136](https://doi.org/10.1111/rssc.12136)
- MAGI, A., TATTINI, L., PIPPUCCI, T., TORRICELLI, F. and BENELLI, M. (2011). Read count approach for DNA copy number variants detection. *Bioinformatics* **28** 470–478.
- MARASS, F., MOULIERE, F., YUAN, K., ROSENFELD, N. and MARKOWETZ, F. (2016). A phylogenetic latent feature model for clonal deconvolution. *Ann. Appl. Stat.* **10** 2377–2404. [MR3592061 https://doi.org/10.1214/16-AOAS986](https://doi.org/10.1214/16-AOAS986)
- MARDIS, E. R. (2008). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9** 387–402. <https://doi.org/10.1146/annurev.genom.9.081307.164359>
- MISALE, S., YAEGER, R., HOBOR, S., SCALA, E., JANAKIRAMAN, M., LISKA, D., VALTORTA, E., SCHIAVO, R., BUSCARINO, M. et al. (2012). Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* **486** 532–536.
- MÜLLER, P., QUINTANA, F. A. and PAGE, G. (2018). Nonparametric Bayesian inference in applications. *Stat. Methods Appl.* **27** 175–206. [MR3807363 https://doi.org/10.1007/s10260-017-0405-z](https://doi.org/10.1007/s10260-017-0405-z)
- NOWELL, P. C. (1976). The clonal evolution of tumor cell populations. *Science* **194** 23–28.
- OESPER, L., MAHMOODY, A. and RAPHAEL, B. J. (2013). THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.* **14** R80. <https://doi.org/10.1186/gb-2013-14-7-r80>
- RADENBAUGH, A. J., MA, S., EWING, A., STUART, J. M., COLLISON, E. A., ZHU, J. and HAUSSLER, D. (2014). RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PLoS ONE* **9** e111516.
- RHEE, J.-K., KIM, K., CHAE, H., EVANS, J., YAN, P., ZHANG, B.-T., GRAY, J., SPELLMAN, P., HUANG, T. H.-M. et al. (2013). Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer. *Nucleic Acids Res.* **41** 8464–8474.
- RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792. [MR1483213 https://doi.org/10.1111/1467-9868.00095](https://doi.org/10.1111/1467-9868.00095)
- ROTH, A., KHATTRA, J., YAP, D., WAN, A., LAKS, E., BIELE, J., HA, G., APARICIO, S., BOUCHARD-CÔTÉ, A. et al. (2014). PyClone: Statistical inference of clonal population structure in cancer. *Nat. Methods* **11** 396–398.
- SCHMIDT, F. and EFFERTH, T. (2016). Tumor heterogeneity, single-cell sequencing, and drug resistance. *Pharmaceuticals (Basel)* **9** 33.
- SCHMITT, M. W., LOEB, L. A. and SALK, J. J. (2016). The influence of subclonal resistance mutations on targeted cancer therapy. *Nature Reviews Clinical Oncology* **13** 335–347.
- SHACKLETON, M., QUINTANA, E., FEARON, E. R. and MORRISON, S. J. (2009). Heterogeneity in cancer: Cancer stem cells versus clonal evolution. *Cell* **138** 822–829.
- 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.

- 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.
- WEST, M. (2003). Bayesian factor regression models in the “large p , small n ” paradigm. In *Bayesian Statistics, 7 (Tenerife, 2002)* 733–742. Oxford Univ. Press, New York. [MR2003537](#)
- WILKERSON, M. D., CABANSKI, C. R., SUN, W., HOADLEY, K. A., WALTER, V., MOSE, L. E., TROESTER, M. A., HAMMERMAN, P. S., PARKER, J. S. et al. (2014). Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Res.* **42** e107.
- 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. [MR3852690](#) <https://doi.org/10.1214/17-AOAS1123>
- ZARE, H., WANG, J., HU, A., WEBER, K., SMITH, J., NICKERSON, D., SONG, C., WITTEN, D., BLAU, C. A. et al. (2014). Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.* **10** e1003703.
- ZENG, L., WARREN, J. L. and ZHAO, H. (2019). Phylogeny-based tumor subclone identification using a Bayesian feature allocation model. *Ann. Appl. Stat.* **13** 1212–1241. [MR3963569](#) <https://doi.org/10.1214/18-AOAS1223>
- ZHOU, T., MÜLLER, P., SENGUPTA, S. and JI, Y. (2019a). PairClone: A Bayesian subclone caller based on mutation pairs. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 705–725. [MR3937470](#) <https://doi.org/10.1111/rssc.12328>
- ZHOU, T., SENGUPTA, S., MÜLLER, P. and JI, Y. (2019b). TreeClone: Reconstruction of tumor subclone phylogeny based on mutation pairs using next generation sequencing data. *Ann. Appl. Stat.* **13** 874–899. [MR3963556](#) <https://doi.org/10.1214/18-AOAS1224>
- ZHOU, T., SENGUPTA, S., MÜLLER, P. and JI, Y. (2020). Supplement to “RNDClone: Tumor subclone reconstruction based on integrating DNA and RNA sequence data.” <https://doi.org/10.1214/20-AOAS1368SUPPA>, <https://doi.org/10.1214/20-AOAS1368SUPPB>.

A BAYESIAN TIME-VARYING EFFECT MODEL FOR BEHAVIORAL MHEALTH DATA

BY MATTHEW D. KOSLOVSKY¹, EMILY T. HÉBERT^{2,*}, MICHAEL S. BUSINELLE^{2,†} AND
MARINA VANNUCCI³

¹Department of Statistics, Colorado State University, matt.koslovsky@colostate.edu

²Oklahoma Tobacco Research Center, University of Oklahoma Health Sciences Center, *Emily-Hebert@ouhsc.edu;

†Michael-Businelle@ouhsc.edu

³Department of Statistics, Rice University, marina@rice.edu

The integration of mobile health (mHealth) devices into behavioral health research has fundamentally changed the way researchers and interventionists are able to collect data as well as deploy and evaluate intervention strategies. In these studies, researchers often collect intensive longitudinal data (ILD) using ecological momentary assessment methods which aim to capture psychological, emotional and environmental factors that may relate to a behavioral outcome in near real time. In order to investigate ILD collected in a novel, smartphone-based smoking cessation study, we propose a Bayesian variable selection approach for time-varying effect models, designed to identify dynamic relations between potential risk factors and smoking behaviors in the critical moments around a quit attempt. We use parameter-expansion and data-augmentation techniques to efficiently explore how the underlying structure of these relations varies over time and across subjects. We achieve deeper insights into these relations by introducing nonparametric priors for regression coefficients that cluster similar effects for risk factors while simultaneously determining their inclusion. Results indicate that our approach is well positioned to help researchers effectively evaluate, design and deliver tailored intervention strategies in the critical moments surrounding a quit attempt.

REFERENCES

- BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. [MR2065192 https://doi.org/10.1214/009053604000000238](https://doi.org/10.1214/009053604000000238)
- BERARDI, V., CARRETERO-GONZÁLEZ, R., BELLETTIERE, J., ADAMS, M. A., HUGHES, S. and HOVELL, M. (2018). A Markov approach for increasing precision in the assessment of data-intensive behavioral interventions. *J. Biomed. Inform.* **85** 93–105.
- BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. (2019). Lasso meets horseshoe: A survey. *Statist. Sci.* **34** 405–427. [MR4017521 https://doi.org/10.1214/19-STS700](https://doi.org/10.1214/19-STS700)
- BOLMAN, C., VERBOON, P., THEWISSEN, V., BOONEN, V., SOONS, K. and JACOBS, N. (2018). Predicting smoking lapses in the first week of quitting: An ecological momentary assessment study. *J. Addict. Med.* **12** 65–71. <https://doi.org/10.1097/ADM.0000000000000365>
- BROOK, D. W., BROOK, J. S., ZHANG, C., WHITEMAN, M., COHEN, P. and FINCH, S. J. (2008). Developmental trajectories of cigarette smoking from adolescence to the early thirties: Personality and behavioral risk factors. *Nicotine Tob. Res.* **10** 1283–1291.
- BROWN, P. J., VANNUCCI, M. and FEARN, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 627–641. [MR1626005 https://doi.org/10.1111/1467-9868.00144](https://doi.org/10.1111/1467-9868.00144)
- BUSINELLE, M. S., MA, P., KENDZOR, D. E., REITZEL, L. R., CHEN, M., LAM, C. Y., BERNSTEIN, I. and WETTER, D. W. (2014). Predicting quit attempts among homeless smokers seeking cessation treatment: An ecological momentary assessment study. *Nicotine Tob. Res.* **16** 1371–1378.
- BUSINELLE, M. S., MA, P., KENDZOR, D. E., FRANK, S. G., VIDRINE, D. J. and WETTER, D. W. (2016). An ecological momentary intervention for smoking cessation: Evaluation of feasibility and effectiveness. *J. Med. Internet Res.* **18** e321.

Key words and phrases. Ecological momentary assessment, mHealth, Pólya-Gamma augmentation, time-varying effect model, variable selection.

- CAI, B. and BANDYOPADHYAY, D. (2017). Bayesian semiparametric variable selection with applications to periodontal data. *Stat. Med.* **36** 2251–2264. [MR3660129](#) <https://doi.org/10.1002/sim.725>
- CANALE, A., LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2017). On the Pitman–Yor process with spike and slab base measure. *Biometrika* **104** 681–697. [MR3694590](#) <https://doi.org/10.1093/biomet/asx041>
- CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics* 73–80.
- CHEN, Z. and DUNSON, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59** 762–769. [MR2025100](#) <https://doi.org/10.1111/j.0006-341X.2003.00089.x>
- CHENG, J., EDWARDS, L. J., MALDONADO-MOLINA, M. M., KOMRO, K. A. and MULLER, K. E. (2010). Real longitudinal data analysis for real people: Building a good enough mixed model. *Stat. Med.* **29** 504–520. [MR2751785](#) <https://doi.org/10.1002/sim.3775>
- CHIPMAN, H. (1996). Bayesian variable selection with related predictors. *Canad. J. Statist.* **24** 17–36. [MR1394738](#) <https://doi.org/10.2307/3315687>
- CURSIO, J. F., MERMELSTEIN, R. J. and HEDEKER, D. (2019). Latent trait shared-parameter mixed models for missing ecological momentary assessment data. *Stat. Med.* **38** 660–673. [MR3902605](#) <https://doi.org/10.1002/sim.7989>
- DAHL, D. B. and MULLER, P. (2017). sdols: Summarizing distributions of latent structures. R Package Version 1.4. 591.
- DE HAAN-RIETDIJK, S., KUPPENS, P., BERGEMAN, C. S., SHEEBER, L. B., ALLEN, N. B. and HAMAKER, E. L. (2017). On the use of mixed Markov models for intensive longitudinal data. *Multivar. Behav. Res.* **52** 747–767. <https://doi.org/10.1080/00273171.2017.1370364>
- DUNSON, D. B., HERRING, A. H. and ENGEL, S. M. (2008). Bayesian selection and clustering of polymorphisms in functionally related genes. *J. Amer. Statist. Assoc.* **103** 534–546. [MR2523991](#) <https://doi.org/10.1198/016214507000000554>
- DZIAK, J. J., LI, R., TAN, X., SHIFFMAN, S. and SHIYKO, M. P. (2015). Modeling intensive longitudinal data with mixtures of nonparametric trajectories and time-varying effects. *Psychol. Methods* **20** 444.
- ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. [MR1340510](#)
- FITZMAURICE, G. M., LAIRD, N. M. and WARE, J. H. (2012). *Applied Longitudinal Analysis*, 2nd ed. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ. [MR2830137](#)
- GEISER, C., BISHOP, J., LOCKHART, G., SHIFFMAN, S. and GRENARD, J. L. (2013). Analyzing latent state-trait and multiple-indicator latent growth curve models as multilevel structural equation models. *Front. Psychol.* **4** 975. <https://doi.org/10.3389/fpsyg.2013.00975>
- GELFAND, A. E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice. Interdiscip. Statist.* 145–161. CRC Press, London. [MR1397969](#)
- GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
- GELMAN, A., GOEGBEUR, Y., TUERLINCKX, F. and VAN MECHELEN, I. (2000). Diagnostic checks for discrete data regression models using posterior predictive simulations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **49** 247–268. [MR1821324](#) <https://doi.org/10.1111/1467-9876.00190>
- GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.
- GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* 339–373.
- HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. [MR1229881](#)
- HERON, K. E. and SMYTH, J. M. (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. *Br. J. Health Psychol.* **15** 1–39. <https://doi.org/10.1348/135910709X466063>
- HUI, F. K. C., MÜLLER, S. and WELSH, A. H. (2017). Hierarchical selection of fixed and random effects in generalized linear mixed models. *Statist. Sinica* **27** 501–518. [MR3674683](#)
- 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>
- JONES, R. H., XU, S. and GRUNWALD, G. K. (2006). Continuous time Markov models for binary longitudinal data. *Biom. J.* **48** 411–419. [MR2240089](#) <https://doi.org/10.1002/bimj.200510224>
- KIM, S., DAHL, D. B. and VANNUCCI, M. (2009). Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Anal.* **4** 707–732. [MR2570085](#) <https://doi.org/10.1214/09-BA426>
- KINNEY, S. K. and DUNSON, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics* **63** 690–698. [MR2395705](#) <https://doi.org/10.1111/j.1541-0420.2007.00771.x>

- KLASNJA, P., HEKLER, E. B., SHIFFMAN, S., BORUVKA, A., ALMIRALL, D., TEWARI, A. and MURPHY, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology* **34** 1220.
- KOSLOVSKY, M. D., HÉBERT, E. T., SWARTZ, M. D., CHAN, W., LEON-NOVELO, L., WILKINSON, A. V., KENDZOR, D. E. and BUSINELLE, M. S. (2018). The time-varying relations between risk factors and smoking before and after a quit attempt. *Nicotine Tob. Res.* **20** 1231–1236.
- KOSLOVSKY, M. D., SWARTZ, M. D., CHAN, W., LEON-NOVELO, L., WILKINSON, A. V., KENDZOR, D. E. and BUSINELLE, M. S. (2018). Bayesian variable selection for multistate Markov models with interval-censored data in an ecological momentary assessment study of smoking cessation. *Biometrics* **74** 636–644. MR3825350 <https://doi.org/10.1111/biom.12792>
- KOSLOVSKY, M. D., HÉBERT, E. T., BUSINELLE, M. S. and VANNUCCI, M. (2020). Supplement to “A Bayesian time-varying effect model for behavioral mHealth data.” <https://doi.org/10.1214/20-AOAS1402SUPPA>, <https://doi.org/10.1214/20-AOAS1402SUPPB>
- KÜRÜM, E., LI, R., SHIFFMAN, S. and YAO, W. (2016). Time-varying coefficient models for joint modeling binary and continuous outcomes in longitudinal data. *Statist. Sinica* **26** 979–1000. MR3559939
- LANG, S. and BREZGER, A. (2004). Bayesian P-splines. *J. Comput. Graph. Statist.* **13** 183–212. MR2044877 <https://doi.org/10.1198/1061860043010>
- LANZA, S. T., VASILENKO, S. A., LIU, X., LI, R. and PIPER, M. E. (2014). Advancing the understanding of craving during smoking cessation attempts: A demonstration of the time-varying effect model. *Nicotine Tob. Res.* **16** S127–S134.
- LI, Y., MÜLLER, P. and LIN, X. (2011). Center-adjusted inference for a nonparametric Bayesian random effect distribution. *Statist. Sinica* **21** 1201–1223. MR2827521 <https://doi.org/10.5705/ss.2009.180>
- LUCKETT, D. J., LABER, E. B., KAHKOSKA, A. R., MAAHS, D. M., MAYER-DAVIS, E. and KOSOROK, M. R. (2020). Estimating dynamic treatment regimes in mobile health using V-learning. *J. Amer. Statist. Assoc.* **115** 692–706. MR4107673 <https://doi.org/10.1080/01621459.2018.1537919>
- MA, S. and SONG, P. X.-K. (2015). Varying index coefficient models. *J. Amer. Statist. Assoc.* **110** 341–356. MR3338507 <https://doi.org/10.1080/01621459.2014.903185>
- MASON, M., MENNIS, J., WAY, T., LANZA, S., RUSSELL, M. and ZAHARAKIS, N. (2015). Time-varying effects of a text-based smoking cessation intervention for urban adolescents. *Drug Alcohol Depend.* **157** 99–105. <https://doi.org/10.1016/j.drugalcdep.2015.10.016>
- MCCARTHY, D. E., EBSSA, L., WITKIEWITZ, K. and SHIFFMAN, S. (2016). Repeated measures latent class analysis of daily smoking in three smoking cessation studies. *Drug Alcohol Depend.* **165** 132–142.
- MCCLURE, J. B., ANDERSON, M. L., BRADLEY, K., AN, L. C. and CATZ, S. L. (2016). Evaluating an adaptive and interactive mHealth smoking cessation and medication adherence program: A randomized pilot feasibility study. *JMIR mHealth uHealth* **4** e94.
- MEILÄ, M. (2003). Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines* 173–187. Springer, Berlin.
- MINAMI, H., YEH, V. M., BOLD, K. W., CHAPMAN, G. B. and MCCARTHY, D. E. (2014). Relations among affect, abstinence motivation and confidence, and daily smoking lapse risk. *Psychol. Addict. Behav.* **28** 376–388. <https://doi.org/10.1037/a0034445>
- MÜLLER, P., QUINTANA, F. A. and ROSNER, G. L. (2007). Semiparametric Bayesian inference for multilevel repeated measurement data. *Biometrics* **63** 280–289, 316. MR2345598 <https://doi.org/10.1111/j.1541-0420.2006.00668.x>
- MÜLLER, S., SCEALY, J. L. and WELSH, A. H. (2013). Model selection in linear mixed models. *Statist. Sci.* **28** 135–167. MR3112403 <https://doi.org/10.1214/12-sts410>
- NAHUM-SHANI, I., SMITH, S. N., SPRING, B. J., COLLINS, L. M., WITKIEWITZ, K., TEWARI, A. and MURPHY, S. A. (2017). Just-in-time adaptive interventions (JITIs) in mobile health: Key components and design principles for ongoing health behavior support. *Ann. Behav. Med.* **52** 446–462.
- NAUGHTON, F., HOPEWELL, S., LATHIA, N., SCHALBROECK, R., BROWN, C., MASCOLO, C., MCEWEN, A. and SUTTON, S. (2016). A context-sensing mobile phone app (Q sense) for smoking cessation: A mixed-methods study. *JMIR mHealth uHealth* **4** e106.
- 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>
- NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.
- PIASECKI, T. M., FIORE, M. C., MCCARTHY, D. E. and BAKER, T. B. (2002). Have we lost our way? The need for dynamic formulations of smoking relapse proneness. *Addiction* **97** 1093–1108.
- PIASECKI, T. M., JORENBY, D. E., SMITH, S. S., FIORE, M. C. and BAKER, T. B. (2003). Smoking withdrawal dynamics: II. Improved tests of withdrawal-relapse relations. *J. Abnorm. Psychol.* **112** 14–27.

- PIASECKI, T. M., TRELA, C. J., HEDEKER, D. and MERMELSTEIN, R. J. (2013). Smoking antecedents: Separating between-and within-person effects of tobacco dependence in a multiwave ecological momentary assessment investigation of adolescent smoking. *Nicotine Tob. Res.* **16** S119–S126.
- POLSON, N. G. and SCOTT, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Stat.* **9** 105.
- POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. [MR3174712](#) <https://doi.org/10.1080/01621459.2013.829001>
- RILEY, W. T., RIVERA, D. E., ATIENZA, A. A., NILSEN, W., ALLISON, S. M. and MERMELSTEIN, R. (2011). Health behavior models in the age of mobile interventions: Are our theories up to the task? *Transl. Behav. Med.* **1** 53–71. <https://doi.org/10.1007/s13142-011-0021-7>
- RIVERA, D. E., PEW, M. D. and COLLINS, L. M. (2007). Using engineering control principles to inform the design of adaptive interventions: A conceptual introduction. *Drug Alcohol Depend.* **88** S31–S40.
- SAVITSKY, T. and VANNUCCI, M. (2010). Spiked Dirichlet process priors for Gaussian process models. *J. Probab. Stat.* **2010** Art. ID 201489, 14. [MR2745498](#) <https://doi.org/10.1155/2010/201489>
- SAVITSKY, T., VANNUCCI, M. and SHA, N. (2011). Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statist. Sci.* **26** 130–149. [MR2849913](#) <https://doi.org/10.1214/11-STS354>
- SCHIEPL, F. (2011). spikeSlabGAM: Bayesian variable selection, model choice and regularization for generalized additive mixed models in R. Preprint. Available at [arXiv:1105.5253](#).
- SCHIEPL, F., FAHRMEIR, L. and KNEIB, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *J. Amer. Statist. Assoc.* **107** 1518–1532. [MR3036413](#) <https://doi.org/10.1080/01621459.2012.737742>
- SELYA, A. S., UPDEGROVE, N., ROSE, J. S., DIERKER, L., TAN, X., HEDEKER, D., LI, R. and MERMELSTEIN, R. J. (2015). Nicotine-dependence-varying effects of smoking events on momentary mood changes among adolescents. *Addict. Behav.* **41** 65–71.
- SHA, N., TADESSE, M. G. and VANNUCCI, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics* **22** 2262–2268.
- SHIFFMAN, S. (2013). Conceptualizing analyses of ecological momentary assessment data. *Nicotine Tob. Res.* **16** S76–S87.
- SHIFFMAN, S., PATY, J. A., GNYS, M., KASSEL, J. A. and HICKCOX, M. (1996). First lapses to smoking: Within-subjects analysis of real-time reports. *J. Consult. Clin. Psychol.* **64** 366–379. <https://doi.org/10.1037/0022-006x.64.2.366>
- SHIFFMAN, S., BALABANIS, M. H., PATY, J. A., ENGBERG, J., GWALTNEY, C. J., LIU, K. S., GNYS, M., HICKCOX, M. and PATON, S. M. (2000). Dynamic effects of self-efficacy on smoking lapse and relapse. *Health Psychology* **19** 315.
- SHIFFMAN, S., GWALTNEY, C. J., BALABANIS, M. H., LIU, K. S., PATY, J. A., KASSEL, J. D., HICKCOX, M. and GNYS, M. (2002). Immediate antecedents of cigarette smoking: An analysis from ecological momentary assessment. *J. Abnorm. Psychol.* **111** 531.
- SHIYKO, M. P., LANZA, S. T., TAN, X., LI, R. and SHIFFMAN, S. (2012). Using the time-varying effect model (TVEM) to examine dynamic associations between negative affect and self confidence on smoking urges: Differences between successful quitters and relapsers. *Prev. Sci.* **13** 288–299.
- SHIYKO, M., NAAB, P., SHIFFMAN, S. and LI, R. (2014). Modeling complexity of EMA data: Time-varying lagged effects of negative affect on smoking urges for subgroups of nicotine addiction. *Nicotine Tob. Res.* **16** S144–S150.
- TAN, X., SHIYKO, M. P., LI, R., LI, Y. and DIERKER, L. (2012). A time-varying effect model for intensive longitudinal data. *Psychol. Methods* **17** 61.
- TIMMS, K. P., RIVERA, D. E., COLLINS, L. M. and PIPER, M. E. (2013). A dynamical systems approach to understanding self-regulation in smoking cessation behavior change. *Nicotine Tob. Res.* **16** S159–S168.
- TRAIL, J. B., COLLINS, L. M., RIVERA, D. E., LI, R., PIPER, M. E. and BAKER, T. B. (2014). Functional data analysis for dynamical system identification of behavioral processes. *Psychol. Methods* **19** 175.
- VAN ERP, S., OBERSKI, D. L. and MULDER, J. (2019). Shrinkage priors for Bayesian penalized regression. *J. Math. Psych.* **89** 31–50. [MR3903921](#) <https://doi.org/10.1016/j.jmp.2018.12.004>
- VASILENKO, S. A., PIPER, M. E., LANZA, S. T., LIU, X., YANG, J. and LI, R. (2014). Time-varying processes involved in smoking lapse in a randomized trial of smoking cessation therapies. *Nicotine Tob. Res.* **16** S135–S143.
- VEHTARI, A., GELMAN, A. and GABRY, J. (2016). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R Package Version 0.1 **6**.

- VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27** 1413–1432. [MR3647105](#) <https://doi.org/10.1007/s11222-016-9696-4>
- WADE, S. and GHAHRAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Anal.* **13** 559–626. [MR3807860](#) <https://doi.org/10.1214/17-BA1073>
- WALLS, T. A. and SCHAFER, J. L., eds. (2006). *Models for Intensive Longitudinal Data* Oxford Univ. Press, Oxford. [MR2334216](#) <https://doi.org/10.1093/acprof:oso/9780195173444.001.0001>
- YANG, M. (2012). Bayesian variable selection for logistic mixed model with nonparametric random effects. *Comput. Statist. Data Anal.* **56** 2663–2674. [MR2915153](#) <https://doi.org/10.1016/j.csda.2011.12.014>

BAYESIAN PROFILING MULTIPLE IMPUTATION FOR MISSING HEMOGLOBIN VALUES IN ELECTRONIC HEALTH RECORDS

BY YAJUAN SI¹, MARI PALTA^{2,*} AND MAUREEN SMITH^{2,†}

¹Survey Research Center, Institute for Social Research, University of Michigan, yajuan@umich.edu

²Department of Population Health Sciences, University of Wisconsin-Madison, *mpalta@wisc.edu; †maureensmith@wisc.edu

Electronic health records (EHRs) are increasingly used for clinical and comparative effectiveness research but suffer from missing data. Motivated by health services research on diabetes care, we seek to increase the quality of EHRs by focusing on missing values of longitudinal glycosylated hemoglobin (A1c), a key risk factor for diabetes complications and adverse events. Under the framework of multiple imputation (MI), we propose an individualized Bayesian latent profiling approach to capture A1c measurement trajectories subject to missingness. The proposed method is applied to EHRs of adult patients with diabetes in a large academic Midwestern health system between 2003 and 2013 and had Medicare A and B coverage. We combine MI inferences to evaluate the association of A1c levels with the incidence of acute adverse health events and examine patient heterogeneity across identified patient profiles. We investigate different missingness mechanisms and perform imputation diagnostics. Our approach is computationally efficient and fits flexible models that provide useful clinical insights.

REFERENCES

- AGENCY FOR HEALTHCARE RESEARCH AND QUALITY (2013). Primary care patients use interactive preventive health record integrated with electronic health record, leading to enhanced provision of preventive services. Available at <https://innovations.ahrq.gov/profiles/primary-care-patients-use-interactive-preventive-health-record-integrated-electronic-health>.
- ADA (2018a). Glycemic targets. *Diabetes Care* **41** 55–64.
- ADA (2018b). Older adults. *Diabetes Care* **41** 119–125.
- AIROLDI, E. M., BLEI, D. M., EROSHOVA, E. A. and FIENBERG, S. E., eds. (2015). *Handbook of Mixed Membership Models and Their Applications*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL. [MR3381023](#)
- ALLEN, C., LECAIRE, T., PALTA, M., DANIELS, K., MEREDITH, M. and D’ALESSIO, D. (2001). Risk factors for frequent and severe hypoglycemia in type 1 diabetes. *Diabetes Care* **24** 1878–1881.
- BOYD, C. M., DARER, J., BOULT, C., FRIED, L. P., BOULT, L. and WU, A. W. (2005). Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases—Implications for pay for performance. *JAMA J. Am. Med. Assoc.* **294** 716–724.
- CARPENTER, J. and KENWARD, M. G. (2013). *Multiple Imputation and Its Application*. Wiley, New York.
- CEBUL, R. D., LOVE, T. E., JAIN, A. K. and HEBERT, C. J. (2011). Electronic health records and quality of diabetes care. *N. Engl. J. Med.* **365** 825–833.
- DANIELS, M. J. and HOGAN, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Monographs on Statistics and Applied Probability **109**. CRC Press, Boca Raton, FL. [MR2459796](#) <https://doi.org/10.1201/9781420011180>
- DAVIS, S. and ALONSO, M. D. (2004). Hypoglycemia as a barrier to glycemic control. *J. Diabetes Complicat.* **18** 60–68.
- DLUHY, R. G. and MCMAHON, G. T. (2008). Intensive glycemic control in the ACCORD and ADVANCE trials. *N. Engl. J. Med.* **358** 2630–2633.
- ELIXHAUSER, A., STEINER, C., HARRIS, D. R. and COFFEY, R. M. (1998). Comorbidity measures for use with administrative data. *Med. Care* **36** 8–27.
- GELFAND, A. E. and DEY, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56** 501–514. [MR1278223](#)

Key words and phrases. Trajectory, latent profile, multiple imputation, posterior predictive check.

- 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., HILL, J., SU, Y.-S., YAJIMA, M., PITTAU, M., GOODRICH, B., SI, Y. and KROPKO, J. (2015). R package mi: Missing Data Imputation and Model Checking. Available at <https://cran.r-project.org/web/packages/mi>.
- GOODMAN, L. A. (2007). On the assignment of individuals to classes. *Sociol. Method.* **37** 1–22.
- 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** Art. ID 16.
- HE, Y., ZASLAVSKY, A. M., LANDRUM, M. B., HARRINGTON, D. P. and CATALANO, P. (2010). Multiple imputation in a large-scale complex survey: A practical guide. *Stat. Methods Med. Res.* **19** 653–670. [MR2744515 https://doi.org/10.1177/0962280208101273](https://doi.org/10.1177/0962280208101273)
- IBRAHIM, J. G. and MOLENBERGHS, G. (2009). Missing data methods in longitudinal studies: A review. *TEST* **18** 1–43. [MR2495958 https://doi.org/10.1007/s11749-009-0138-x](https://doi.org/10.1007/s11749-009-0138-x)
- IBRAHIM, J. G., CHEN, M.-H., LIPSITZ, S. R. and HERRING, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *J. Amer. Statist. Assoc.* **100** 332–346. [MR2166072 https://doi.org/10.1198/016214504000001844](https://doi.org/10.1198/016214504000001844)
- JACKSON, G. L., EDELMAN, D. and WEINBERGER, M. (2006). Simultaneous control of intermediate diabetes outcomes among veterans affairs primary care patients. *J. Gen. Intern. Med.* **21** 1050–1056.
- KENNEDY, R. L., HENRY, J., CHAPMAN, A., NAYAR, R., GRANT, P. and MORRIS, A. D. (2002). Accidents in patients with insulin-treated diabetes: Increased risk of low-impact falls but not motor vehicle crashes—A prospective register-based study. *J. Trauma Inj. Infect. Crit. Care* **52** 660–666.
- KING, G., HONAKER, J., JOSEPH, A. and SCHEVE, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *Am. Polit. Sci. Rev.* **95** 49–69.
- LI, F.-R., ZHANG, X.-R., ZHONG, W.-F., LI, Z.-H., GAO, X., KRAUS, V. B., LV, Y.-B., ZOU, M.-C., CHEN, G.-C. et al. (2019). Glycated hemoglobin and all-cause and cause-specific mortality among adults with and without diabetes. *J. Clin. Endocrinol. Metab.*
- LIN, H., MCCULLOCH, C. E. and ROSENHECK, R. A. (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics* **60** 295–305. [MR2066263 https://doi.org/10.1111/j.0006-341X.2004.00173.x](https://doi.org/10.1111/j.0006-341X.2004.00173.x)
- LITTLE, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *J. Amer. Statist. Assoc.* **90** 1112–1121. [MR1354029](https://doi.org/10.1080/01621459.2013.829001)
- LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. [MR1925014 https://doi.org/10.1002/9781119013563](https://doi.org/10.1002/9781119013563)
- MARTIN, C. L., ALBERS, J., HERMAN, W. H., CLEARY, P., WABERSKI, B., GREENE, D. A., STEVENS, M. J. and FELDMAN, E. L. (2006). Neuropathy among the diabetes control and complications trial cohort 8 years after trial completion. *Diabetes Care* **29** 340–344.
- MCCULLOCH, C. E., NEUHAUS, J. M. and OLIN, R. L. (2016). Biased and unbiased estimation in longitudinal studies with informative visit processes. *Biometrics* **72** 1315–1324. [MR3591616 https://doi.org/10.1111/biom.12501](https://doi.org/10.1111/biom.12501)
- McFARLANE, S. I., JACOB, S. J., WINER, N., KAUR, J., CASTRO, J. P., WUI, M. A., GLIWA, A., VON GIZYCKI, H. and SOWERS, J. R. (2002). Control of cardiovascular risk factors in patients with diabetes and hypertension at urban academic medical centers. *Diabetes Care* **25** 718–723.
- MENG, X.-L. (1994). Posterior predictive *p*-values. *Ann. Statist.* **22** 1142–1160. [MR1311969 https://doi.org/10.1214/aos/1176325622](https://doi.org/10.1214/aos/1176325622)
- NATIONAL RESEARCH COUNCIL (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. The National Academies Press, Washington, DC.
- NIEFELD, M. R., BRAUNSTEIN, J. B., WU, A. W., SAUDEK, C. D., WELLER, W. E. and ANDERSON, G. F. (2003). Preventable hospitalization among elderly medicare beneficiaries with type 2 diabetes. *Diabetes Care* **26** 1344–1349.
- POGACH, L. M., TIWARI, A., MANEY, M., RAJAN, M., MILLER, D. R. and ARON, D. (2007). Should mitigating comorbidities be considered in assessing healthcare plan performance in achieving optimal glycemic control? *Am. J. Manag. Care* **13** 133–140.
- POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. [MR3174712 https://doi.org/10.1080/01621459.2013.829001](https://doi.org/10.1080/01621459.2013.829001)
- QASEERN, A., VIJAN, S., SNOW, V., CROSS, J. T., WEISS, K. B. and OWENS, D. K. (2007). Glycemic control and type 2 diabetes mellitus: The optimal hemoglobin a1c targets—A guidance statement from the American College of Physicians. *Ann. Intern. Med.* **147** 417–422.

- RAGHUNATHAN, T. E., LEPKOWSKI, J. M., VAN HOEWYK, J. and SOLENBERGER, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.* **27** 85–96.
- ROBINS, J. M., HERNÁN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.
- RODRÍGUEZ, A. and DUNSON, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal.* **6** 145–177. [MR2781811 https://doi.org/10.1214/11-BA605](https://doi.org/10.1214/11-BA605)
- RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. [MR0899519 https://doi.org/10.1002/9780470316696](https://doi.org/10.1002/9780470316696)
- SAS INSTITUTE INC. (2017). SAS/STAT 9.4 User’s Guide.
- 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](https://doi.org/10.1201/9781439821862)
- SCHAFER, J. L. (2016). R package pan: Multiple imputation for multivariate panel or clustered data. Available at <https://cran.r-project.org/web/packages/pan/>.
- SCHECH, S., GRAHAM, D., STAFFA, J., ANDRADE, S. E., LA GRENADE, L., BURGESS, M., BLOUGH, D., STERGACHIS, A., CHAN, K. A. et al. (2007). Risk factors for statin-associated rhabdomyolysis. *Pharmacoepidemiol. Drug Saf.* **16** 352–358.
- SHORR, R. I., RAY, W. A., DAUGHERTY, J. R. and GRIFFIN, M. R. (1997). Incidence and risk factors for serious hypoglycemia in older persons using insulin or sulfonylureas. *Arch. Intern. Med.* **157** 1681–1686.
- SI, Y., PALTA, M. and SMITH, M. (2020). Supplement to “Bayesian profiling multiple imputation for missing hemoglobin values in electronic health records.” <https://doi.org/10.1214/20-AOAS1378SUPP>
- SI, Y. and REITER, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *J. Educ. Behav. Stat.* **38** 499–521.
- SI, Y., REITER, J. P. and HILLYGUS, D. S. (2016). Bayesian latent pattern mixture models for handling attrition in panel studies with refreshment samples. *Ann. Appl. Stat.* **10** 118–143. [MR3480490 https://doi.org/10.1214/15-AOAS876](https://doi.org/10.1214/15-AOAS876)
- SUN, B. and TCHETGEN TCHETGEN, E. J. (2018). On inverse probability weighting for nonmonotone missing at random data. *J. Amer. Statist. Assoc.* **113** 369–379. [MR3803471 https://doi.org/10.1080/01621459.2016.1256814](https://doi.org/10.1080/01621459.2016.1256814)
- THE CENTERS FOR DISEASE CONTROL AND PREVENTION (2017). National diabetes statistics report. Available at <https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf>.
- U.S. CENTER FOR MEDICARE & MEDICAID SERVICES (2012). Electronic health records. Available at <https://www.cms.gov/Medicare/E-Health/EHealthRecords/index.html>.
- VAN BUUREN, S. and OUDSHOORN, C. (1999). Flexible multivariate imputation by MICE. Technical report, TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054, Leiden.
- VERMUNT, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Polit. Anal.* **18** 450–468.
- WU, M. C. and CARROLL, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44** 175–188. [MR0931633 https://doi.org/10.2307/2531905](https://doi.org/10.2307/2531905)
- ZELDOW, B., FLORY, J., STEPHENS-SHIELDS, A., RAEBEL, M. and ROY, J. (2018). Outcome identification in electronic health records using predictions from an enriched Dirichlet process mixture. Available at [arXiv:1806.02411](https://arxiv.org/abs/1806.02411).

BAYESIAN INFERENCE FOR MULTISTRAIN EPIDEMICS WITH APPLICATION TO *ESCHERICHIA COLI* O157:H7 IN FEEDLOT CATTLE

BY PANAYIOTA TOULOUPOU^{1,*}, BÄRBEL FINKENSTÄDT^{1,†}, THOMAS E. BESSER²,
NIGEL P. FRENCH³ AND SIMON E. F. SPENCER^{1,‡}

¹Department of Statistics, Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research, University of Warwick, *P.Touloupou.1@warwick.ac.uk; †B.F.Finkenstadt@warwick.ac.uk; ‡S.E.F.Spencer@warwick.ac.uk

²Department Veterinary Microbiology and Pathology, Washington State University, tbesser@wsu.edu

³The New Zealand Food Safety Science and Research Centre, School of Veterinary Science, Massey University, N.P.French@massey.ac.nz

For most pathogens, testing procedures can be used to distinguish between different strains with which individuals are infected. Due to the growing availability of such data, multistrain models have increased in popularity over the past few years. Quantifying the interactions between different strains of a pathogen is crucial in order to obtain a more complete understanding of the transmission process, but statistical methods for this type of problem are still in the early stages of development. Motivated by this demand, we construct a stochastic epidemic model that incorporates additional strain information and propose a statistical algorithm for efficient inference. The model improves upon existing methods in the sense that it allows for both imperfect diagnostic test sensitivities and strain misclassification. Extensive simulation studies were conducted in order to assess the performance of our method, while the utility of the developed methodology is demonstrated on data obtained from a longitudinal study of *Escherichia coli* O157:H7 strains in feedlot cattle.

REFERENCES

- ANDERSON, R. M. and MAY, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, London.
- AURANEN, K., ARJAS, E., LEINO, T. and TAKALA, A. K. (2000). Transmission of pneumococcal carriage in families: A latent Markov process model for binary longitudinal data. *J. Amer. Statist. Assoc.* **95** 1044–1053. [MR1821713 https://doi.org/10.2307/2669741](https://doi.org/10.2307/2669741)
- BECKER, N. G. (1989). *Analysis of Infectious Disease Data. Monographs on Statistics and Applied Probability*. CRC Press, London. [MR1014889](https://doi.org/10.1080/00036818908839553)
- BECKER, N. G. (1997). Uses of the EM algorithm in the analysis of data on HIV/AIDS and other infectious diseases. *Stat. Methods Med. Res.* **6** 24–37. <https://doi.org/10.1177/096228029700600103>
- BECKER, N. G. and BRITTON, T. (1999). Statistical studies of infectious disease incidence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 287–307. [MR1680342 https://doi.org/10.1111/1467-9868.00177](https://doi.org/10.1111/1467-9868.00177)
- BECKER, N. G. and HASOFER, A. M. (1997). Estimation in epidemics with incomplete observations. *J. Roy. Statist. Soc. Ser. B* **59** 415–429. [MR1440589 https://doi.org/10.1111/1467-9868.00076](https://doi.org/10.1111/1467-9868.00076)
- CARTER, C. K. and KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81** 541–553. [MR1311096 https://doi.org/10.1093/biomet/81.3.541](https://doi.org/10.1093/biomet/81.3.541)
- CAUCHEMEZ, S., TEMIME, L., VALLERON, A.-J., VARON, E., THOMAS, G., GUILLEMOT, D. and BOËLLE, P.-Y. (2006). *S. pneumoniae* transmission according to inclusion in conjugate vaccines: Bayesian analysis of a longitudinal follow-up in schools. *BMC Infect. Dis.* **6** 1–10.
- CLANCY, D. and O'NEILL, P. D. (2007). Exact Bayesian inference and model selection for stochastic models of epidemics among a community of households. *Scand. J. Stat.* **34** 259–274. [MR2346639 https://doi.org/10.1111/j.1467-9469.2006.00522.x](https://doi.org/10.1111/j.1467-9469.2006.00522.x)
- COBBOLD, R. N., HANCOCK, D. D., RICE, D. H., BERG, J., STILBORN, R., HOVDE, C. J. and BESSER, T. E. (2007). Rectoanal junction colonization of feedlot cattle by *Escherichia coli* O157:H7 and its association with supershedders and excretion dynamics. *Appl. Environ. Microbiol.* **73** 1563–1568.

- DALEY, D. J. and GANI, J. (1999). *Epidemic Modelling: An Introduction*. Cambridge Studies in Mathematical Biology **15**. Cambridge Univ. Press, Cambridge. [MR1688203](#) <https://doi.org/10.1017/CBO9780511608834>
- DIEKMANN, O., HEESTERBEEK, H. and BRITTON, T. (2013). *Mathematical Tools for Understanding Infectious Disease Dynamics*. Princeton Series in Theoretical and Computational Biology. Princeton Univ. Press, Princeton, NJ. [MR3015083](#)
- DONG, W., PENTLAND, A. and HELLER, K. A. (2012). Graph-coupled HMMs for modeling the spread of infection. Preprint. Available at [arXiv:1210.4864](#).
- ERÄSTÖ, P., HOTI, F. and AURANEN, K. (2012). Modeling transmission of multitype infectious agents: Application to carriage of *Streptococcus pneumoniae*. *Stat. Med.* **31** 1450–1463. [MR2931087](#) <https://doi.org/10.1002/sim.4487>
- FEARNHEAD, P. and MELIGKOTSIDOU, L. (2004). Exact filtering for partially observed continuous time models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 771–789. [MR2088781](#) <https://doi.org/10.1111/j.1467-9868.2004.05561.x>
- GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.
- GIBSON, G. J. and RENSHAW, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *Math. Med. Biol.* **15** 19–40.
- IONIDES, E. L., BRETÖ, C. and KING, A. A. (2006). Inference for nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* **103** 18438–18443.
- IONIDES, E. L., NGUYEN, D., ATCHADÉ, Y., STOEV, S. and KING, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proc. Natl. Acad. Sci. USA* **112** 719–724. [MR3311541](#) <https://doi.org/10.1073/pnas.1410597112>
- JEFFREYS, H. (1961). *Theory of Probability*. Third Edition. Clarendon Press, Oxford. [MR0187257](#)
- JEWELL, C. P., KYPRAIOS, T., NEAL, P. and ROBERTS, G. O. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Anal.* **4** 465–496. [MR2551042](#) <https://doi.org/10.1214/09-BA417>
- KELLY, D. and ATWOOD, C. (2011). Finding a minimally informative Dirichlet prior distribution using least squares. *Reliab. Eng. Syst. Saf.* **96** 398–402.
- KYPRAIOS, T., NEAL, P. and PRANGLE, D. (2017). A tutorial introduction to Bayesian inference for stochastic epidemic models using approximate Bayesian computation. *Math. Biosci.* **287** 42–53. [MR3634152](#) <https://doi.org/10.1016/j.mbs.2016.07.001>
- KYPRAIOS, T. and O’NEILL, P. D. (2018). Bayesian nonparametrics for stochastic epidemic models. *Statist. Sci.* **33** 44–56. [MR3757503](#) <https://doi.org/10.1214/17-STS617>
- KYPRAIOS, T., O’NEILL, P. D., HUANG, S. S., RIFAS-SHIMAN, S. L. and COOPER, B. S. (2010). Assessing the role of undetected colonization and isolation precautions in reducing Methicillin-Resistant *Staphylococcus aureus* transmission in intensive care units. *BMC Infect. Dis.* **10** 1–10.
- LORD, C., BARNARD, B., DAY, K., HARGROVE, J., MCNAMARA, J., PAUL, R., TRENHOLME, K. and WOOLHOUSE, M. (1999). Aggregation and distribution of strains in microparasites. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **354** 799–807.
- MCKINLEY, T., COOK, A. R. and DEARDON, R. (2009). Inference in epidemic models without likelihoods. *Int. J. Biostat.* **5** Art. 24, 39. [MR2533810](#) <https://doi.org/10.2202/1557-4679.1171>
- MCKINLEY, T. J., ROSS, J. V., DEARDON, R. and COOK, A. R. (2014). Simulation-based Bayesian inference for epidemic models. *Comput. Statist. Data Anal.* **71** 434–447. [MR3131981](#) <https://doi.org/10.1016/j.csda.2012.12.012>
- MELEGARO, A., CHOI, Y., PEBODY, R. and GAY, N. (2007). *Pneumococcal* carriage in United Kingdom families: Estimating serotype-specific transmission parameters from longitudinal data. *Am. J. Epidemiol.* **166** 228–235.
- MESSINA, J. P., BRADY, O. J., SCOTT, T. W., ZOU, C., PIGOTT, D. M., DUDA, K. A., BHATT, S., KATZELNICK, L., HOWES, R. E. et al. (2014). Global spread of Dengue virus types: Mapping the 70 year history. *Trends Microbiol.* **22** 138–146.
- MORTON, A. and FINKENSTÄDT, B. F. (2005). Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 575–594. [MR2137255](#) <https://doi.org/10.1111/j.1467-9876.2005.05366.x>
- 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](#)
- NEAL, P. (2012). Efficient likelihood-free Bayesian computation for household epidemics. *Stat. Comput.* **22** 1239–1256. [MR2992297](#) <https://doi.org/10.1007/s11222-010-9216-x>
- NORMILE, D. (2013). Tropical medicine. Surprising new Dengue virus throws a spanner in disease control efforts. *Science* **342** 415. [https://doi.org/10.1126/science.342.6157.415](#)

- NUMMINEN, E., CHENG, L., GYLLENBERG, M. and CORANDER, J. (2013). Estimating the transmission dynamics of *Streptococcus pneumoniae* from strain prevalence data. *Biometrics* **69** 748–757. MR3106603 <https://doi.org/10.1111/biom.12040>
- O’NEILL, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Math. Biosci.* **180** 103–114. MR1950750 [https://doi.org/10.1016/S0025-5564\(02\)00109-8](https://doi.org/10.1016/S0025-5564(02)00109-8)
- O’NEILL, P. D. and BECKER, N. G. (2001). Inference for an epidemic when susceptibility varies. *Biostatistics* **2** 99–108. <https://doi.org/10.1093/biostatistics/2.1.99>
- O’NEILL, P. D. and ROBERTS, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *J. Roy. Statist. Soc. Ser. A* **162** 121–129.
- R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- RIDA, W. N. (1991). Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. *J. Roy. Statist. Soc. Ser. B* **53** 269–283. MR1094286
- SMITH, T. and VOUNATSOU, P. (2003). Estimation of infection and recovery rates for highly polymorphic parasites when detectability is imperfect, using hidden Markov models. *Stat. Med.* **22** 1709–1724.
- SPENCER, S. E. F., BESSER, T. E., COBBOLD, R. N. and FRENCH, N. P. (2015). “Super” or just “above average”? Supershedders and the transmission of *Escherichia coli* O157:H7 among feedlot cattle. *J. R. Soc. Interface* **12** 0446. <https://doi.org/10.1098/rsif.2015.0446>
- STREFTARIS, G. and GIBSON, G. J. (2004). Bayesian inference for stochastic epidemics in closed populations. *Stat. Model.* **4** 63–75. MR2037814 <https://doi.org/10.1191/1471082X04st0650a>
- TENOVER, F., ARBEIT, R., GOERING, R., MURRAY, B., PERSING, D., PFALLER, M. and WEINSTEIN, R. (1997). How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: A review for healthcare epidemiologists. *Infection Control and Hospital Epidemiology* **18** 426–439.
- TONI, T., WELCH, D., STRELKOWA, N., IPSEN, A. and STUMPF, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6** 187–202.
- TOULOUPOU, P. (2016). Bayesian inference and model selection for partially observed stochastic epidemics Ph.D. thesis Univ. Warwick.
- TOULOUPOU, P., FINKENSTÄDT, B. and SPENCER, S. E. F. (2020). Scalable Bayesian inference for coupled hidden Markov and semi-Markov models. *J. Comput. Graph. Statist.* **29** 238–249. MR4116038 <https://doi.org/10.1080/10618600.2019.1654880>
- TOULOUPOU, P. and SPENCER, S. E. F. (2020). epiPOMS: Bayesian inference for partially observed multi-strain epidemics. R package, version 0.1.0.
- TOULOUPOU, P., FINKENSTÄDT, B., BESSER, T. E., FRENCH, N. P. and SPENCER, S. E. F. (2020a). Supplement to “Bayesian inference for multistrain epidemics with application to *Escherichia coli* O157:H7 in feedlot cattle.” <https://doi.org/10.1214/20-AOAS1366SUPPA>.
- TOULOUPOU, P., FINKENSTÄDT, B., BESSER, T. E., FRENCH, N. P. and SPENCER, S. E. F. (2020b). Source code for “Bayesian inference for multistrain epidemics with application to *Escherichia coli* O157:H7 in feedlot cattle.” <https://doi.org/10.1214/20-AOAS1366SUPPB>.
- WORBY, C. J., O’NEILL, P. D., KYPRAIOS, T., ROBOTHAM, J. V., DE ANGELIS, D., CARTWRIGHT, E. J. P., PEACOCK, S. J. and COOPER, B. S. (2016). Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *Ann. Appl. Stat.* **10** 395–417. MR3480501 <https://doi.org/10.1214/15-AOAS898>
- XU, X., KYPRAIOS, T. and O’NEILL, P. D. (2016). Bayesian non-parametric inference for stochastic epidemic models using Gaussian processes. *Biostatistics* **17** 619–633. MR3604269 <https://doi.org/10.1093/biostatistics/kxw011>

INTEGRATIVE STATISTICAL METHODS FOR EXPOSURE MIXTURES AND HEALTH

BY BRIAN J. REICH¹, YAWEN GUAN², DENIS FOURCHES³, JOSHUA L. WARREN⁴,
STEFANIE E. SARNAT⁵ AND HOWARD H. CHANG⁶

¹Department of Statistics, North Carolina State University, bjreich@ncsu.edu

²Department of Statistics, University of Nebraska, yguan12@unl.edu

³Department of Chemistry, North Carolina State University, dfourch@ncsu.edu

⁴Department of Biostatistics, Yale University, joshua.warren@yale.edu

⁵Department of Environmental Health, Emory University, sebelt@emory.edu

⁶Department of Biostatistics and Bioinformatics, Emory University, howard.chang@emory.edu

Humans are concurrently exposed to chemically, structurally and toxicologically diverse chemicals. A critical challenge for environmental epidemiology is to quantify the risk of adverse health outcomes resulting from exposures to such chemical mixtures and to identify which mixture constituents may be driving etiologic associations. A variety of statistical methods have been proposed to address these critical research questions. However, they generally rely solely on measured exposure and health data available within a specific study. Advancements in understanding of the role of mixtures on human health impacts may be better achieved through the utilization of external data and knowledge from multiple disciplines with innovative statistical tools. In this paper we develop new methods for health analyses that incorporate auxiliary information about the chemicals in a mixture, such as physico-chemical, structural and/or toxicological data. We expect that the constituents identified using auxiliary information will be more biologically meaningful than those identified by methods that solely utilize observed correlations between measured exposure. We develop flexible Bayesian models by specifying prior distributions for the exposures and their effects that include auxiliary information and examine this idea over a spectrum of analyses from regression to factor analysis. The methods are applied to study the effects of volatile organic compounds on emergency room visits in Atlanta. We find that including cheminformatic information about the exposure variables improves prediction and provides a more interpretable model for emergency room visits for respiratory diseases.

REFERENCES

- ANTONELLI, J., MAZUMDAR, M., BELLINGER, D., CHRISTIANI, D., WRIGHT, R. and COULL, B. (2020). Estimating the health effects of environmental mixtures using Bayesian semiparametric regression and sparsity inducing priors. *Ann. Appl. Stat.* **14** 257–275. MR4085093 <https://doi.org/10.1214/19-AOAS1307>
- ATKINSON, R. W., MILLS, I. C., WALTON, H. A. and ANDERSON, H. R. (2015). Fine particle components and health—a systematic review and meta-analysis of epidemiological time series studies of daily mortality and hospital admissions. *J. Expo. Sci. Environ. Epidemiol.* **25** 208–214. <https://doi.org/10.1038/jes.2014.63>
- BILLIONNET, C., SHERRILL, D., ANNESI-MAESANO, I. and GERIE STUDY (2012). Estimating the health effects of exposure to multi-pollutant mixture. *Ann. Epidemiol.* **22** 126–141.
- BOBB, J. F., VALERI, L., CLAUS HENN, B., CHRISTIANI, D. C., WRIGHT, R. O., MAZUMDAR, M., GODLESKI, J. J. and COULL, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* **16** 493–508. MR3365442 <https://doi.org/10.1093/biostatistics/kxu058>
- CARRICO, C., GENNINGS, C., WHEELER, D. C. and FACTOR-LITVAK, P. (2015). Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *J. Agric. Biol. Environ. Stat.* **20** 100–120. MR3334469 <https://doi.org/10.1007/s13253-014-0180-3>

Key words and phrases. Cheminformatics, collinearity, factor analysis, principal components, stochastic search, variable selection.

- CHEN, R., YIN, P., MENG, X., LIU, C., WANG, L., XU, X., ROSS, J. A., TSE, L. A., ZHAO, Z. et al. (2017). Fine particulate air pollution and daily mortality. A nationwide analysis in 272 Chinese cities. *Am. J. Respir. Crit. Care Med.* **196** 73–81. <https://doi.org/10.1164/rccm.201609-1862OC>
- NATIONAL RESEARCH COUNCIL AND OTHERS (2004a). *Air Quality Management in the United States*. National Academies Press, Washington, DC.
- NATIONAL RESEARCH COUNCIL AND OTHERS (2004b). *Research Priorities for Airborne Particulate Matter: IV. Continuing Research Progress* **4**. National Academies Press, Washington, DC.
- DOMINICI, F., WANG, C., CRAINICEANU, C. and PARMIGIANI, G. (2008). Model selection and health effect estimation in environmental epidemiology. *Epidemiology* **19** 558–560.
- DOMINICI, F., PENG, R. D., BARR, C. D. and BELL, M. L. (2010). Protecting human health from air pollution: Shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology* **21** 187.
- DRAGON (2019). Dragon. Available at http://www.taletе.mi.it/products/dragon_description.htm.
- FANG, X., FANG, B., WANG, C., XIA, T., BOTTAI, M., FANG, F. and CAO, Y. (2019). Comparison of frequentist and Bayesian generalized additive models for assessing the association between daily exposure to fine particles and respiratory mortality: A simulation study. *Int. J. Environ. Res. Public Health* **16** 746.
- FILER, D. L., KOTHIYA, P., SETZER, R. W., JUDSON, R. S. and MARTIN, M. T. (2016). tcpl: The ToxCast pipeline for high-throughput screening data. *Bioinformatics* **33** 618–620.
- FOURCHES, D., MURATOV, E. and TROPSHA, A. (2010). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* **50** 1189–1204.
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. [MR2221284 https://doi.org/10.1214/06-BA117A](https://doi.org/10.1214/06-BA117A)
- GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.
- GHOSH, J. and GHATTAS, A. E. (2015). Bayesian variable selection under collinearity. *Amer. Statist.* **69** 165–173. [MR3391636 https://doi.org/10.1080/00031305.2015.1031827](https://doi.org/10.1080/00031305.2015.1031827)
- GUHA, R. (2007). Chemical informatics functionality in R. *J. Stat. Softw.* **18**.
- HANSEN, D. A., EDGERTON, E., HARTSELL, B., JANSEN, J., BURGE, H., KOUTRAKIS, P., ROGERS, C., SUH, H., CHOW, J. et al. (2006). Air quality measurements for the aerosol research and inhalation epidemiology study. *J. Air Waste Manage. Assoc.* **56** 1445–1458.
- KWON, D., LANDI, M. T., VANNUCCI, M., ISSAQ, H. J., PRIETO, D. and PFEIFFER, R. M. (2011). An efficient stochastic search for Bayesian variable selection with high-dimensional correlated predictors. *Comput. Statist. Data Anal.* **55** 2807–2818. [MR2811867 https://doi.org/10.1016/j.csda.2011.04.019](https://doi.org/10.1016/j.csda.2011.04.019)
- LANDRUM, G. (2019). RDKit: Open-source cheminformatics. Available at <http://www.rdkit.org>.
- LI, C. and LI, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl. Stat.* **4** 1498–1516. [MR2758338 https://doi.org/10.1214/10-AOAS332](https://doi.org/10.1214/10-AOAS332)
- LI, Z., LIU, Z., ZHONG, W., HUANG, M., WU, N., XIE, Y., DAI, Z. and ZOU, X. (2016). Large-scale identification of human protein function using topological features of interaction network. *Sci. Rep.* **6** 37179.
- MACLEHOSE, R. F., DUNSON, D. B., HERRING, A. H. and HOPPIN, J. A. (2015). Bayesian methods for highly correlated exposure data. *Biostatistics* **16** 493–508.
- MOLITER, J., PAPATHOMAS, M., JERRETT, M. and RICHARDSON, S. (2010). Bayesian profile regression with an application to the National Survey of Children’s Health. *Biostatistics* **11** 484–498.
- MORTIMER, K., NEUGEBAUER, R., LURMANN, F., ALCORN, S., BALMES, J. and TAGER, I. (2008). Air pollution and pulmonary function in asthmatic children: Effects of prenatal and lifetime exposures. *Epidemiology* **19** 550–557.
- O’HARA, R. B. and SILLANPÄÄ, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Anal.* **4** 85–117. [MR2486240 https://doi.org/10.1214/09-BA403](https://doi.org/10.1214/09-BA403)
- PEARCE, J. L., WALLER, L. A., SARNAT, S. E., CHANG, H. H., KLEIN, M., MULHOLLAND, J. A. and TOLBERT, P. E. (2016). Characterizing the spatial distribution of multiple pollutants and populations at risk in Atlanta, Georgia. *Spat. Spatio-Tempor. Epidemiol.* **18** 13–23.
- REICH, B. J., GUAN, Y., FOURCHES, D., WARREN, J. L., SARNET, S. E. and CHANG, H. H. (2020). Supplement to “Integrative statistical methods for exposure mixtures and health.” <https://doi.org/10.1214/20-AOAS1364SUPP>
- SABANÉS BOVÉ, D., HELD, L. and KAUERMANN, G. (2015). Objective Bayesian model selection in generalized additive models with penalized splines. *J. Comput. Graph. Statist.* **24** 394–415. [MR3357387 https://doi.org/10.1080/10618600.2014.912136](https://doi.org/10.1080/10618600.2014.912136)
- SHAH, A. S. V., LEE, K. K., MCALLISTER, D. A., HUNTER, A., NAIR, H., WHITELEY, W., LANGRISH, J. P., NEWBY, D. E. and MILLS, N. L. (2015). Short term exposure to air pollution and stroke: Systematic review and meta-analysis. *BMJ* **350** h1295. <https://doi.org/10.1136/bmj.h1295>
- SINISI, S. E. and VAN DER LAAN, M. J. (2004). Deletion/substitution/addition algorithm in learning with applications in genomics. *Stat. Appl. Genet. Mol. Biol.* **3** 18. [MR2101467 https://doi.org/10.2202/1544-6115.1069](https://doi.org/10.2202/1544-6115.1069)

- STIEB, D. M., CHEN, L., ESHOUL, M. and JUDEK, S. (2012). Ambient air pollution, birth weight and preterm birth: A systematic review and meta-analysis. *Environ. Res.* **117** 100–111.
- STRICKLAND, M. J., DARROW, L. A., KLEIN, M., FLANDERS, W. D., SARNAT, J. A., WALLER, L. A., SARNAT, S. E., MULHOLLAND, J. A. and TOLBERT, P. E. (2010). Short-term associations between ambient air pollutants and pediatric asthma emergency department visits. *Am. J. Respir. Crit. Care Med.* **2010** 307–316.
- SUH, H. H., ZANOBBETTI, A., SCHWARTZ, J. and COULL, B. A. (2011). Chemical properties of air pollutants and cause-specific hospital admissions among the elderly in Atlanta, Georgia. *Environ. Health Perspect.* **119** 1421–1428.
- SUN, Z., TAO, Y., LI, S., FERGUSON, K. K., MEEKER, J. D., PARK, S. K., BATTERMAN, S. A. and MUKHERJEE, B. (2013). Statistical strategies for constructing health risk models with multiple pollutants and their interactions: Possible choices and comparisons. *Environ. Health* **12** 85.
- VRIJHEID, M., MARTINEZ, D., MANZANARES, S., DADVAND, P., SCHEMBARI, A., RANKIN, J. and NIEUWENHUIJSEN, M. (2011). Ambient air pollution and risk of congenital anomalies: A systematic review and meta-analysis. *Environ. Health Perspect.* **119** 598–606. <https://doi.org/10.1289/ehp.1002946>
- WEI, R., REICH, B. J., HOPPIN, J. A. and GHOSAL, S. (2020). Sparse Bayesian additive nonparametric regression with application to health effects of pesticides mixtures. *Statist. Sinica* **30** 55–79.
- YE, D., KLEIN, M., CHANG, H. H., SARNAT, J. A., MULHOLLAND, J. A., EDGERTON, E. S., WINQUIST, A., TOLBERT, P. E. and SARNAT, S. E. (2017). Estimating acute cardiorespiratory effects of ambient volatile organic compounds. *Epidemiology* **28** 197–206. <https://doi.org/10.1097/EDE.0000000000000607>
- YE, D., KLEIN, M., MULHOLLAND, J. A., RUSSELL, A. G., WEBER, R., EDGERTON, E. S., CHANG, H. H., SARNAT, J. A., TOLBERT, P. E. et al. (2018). Estimating acute cardiovascular effects of ambient PM. *Environ. Health Perspect.* **126** 027007. <https://doi.org/10.1289/EHP2182>
- ZHENG, X., DING, H., JIANG, L., CHEN, S., ZHENG, J., QIU, M., ZHOU, Y., CHEN, Q. and GUAN, W. (2015). Association between air pollutants and asthma emergency room visits and hospital admissions in time series studies: A systematic review and meta-analysis. *PLoS ONE* **10** e0138146. <https://doi.org/10.1371/journal.pone.0138146>

INCORPORATING COMPOSITIONAL HETEROGENEITY INTO LIE MARKOV MODELS FOR PHYLOGENETIC INFERENCE

BY NAOMI E. HANNAFORD^{1,*}, SARAH E. HEAPS^{1,†}, TOM M. W. NYE^{1,‡},
TOM A. WILLIAMS² AND T. MARTIN EMBLEY³

¹*School of Mathematics, Statistics and Physics, Newcastle University, *n.e.hannaford@ncl.ac.uk; †sarah.heaps@ncl.ac.uk; ‡tom.nye@ncl.ac.uk*

²*School of Biological Sciences, University of Bristol, tom.a.williams@bristol.ac.uk*

³*Institute for Cell and Molecular Biosciences, Newcastle University, martin.embley@ncl.ac.uk*

Phylogenetics uses alignments of molecular sequence data to learn about evolutionary trees. Substitutions in sequences are modelled through a continuous-time Markov process, characterised by an instantaneous rate matrix, which standard models assume is time-reversible and stationary. These assumptions are biologically questionable and induce a likelihood function which is invariant to a tree's root position. This hampers inference because a tree's biological interpretation depends critically on where it is rooted. Relaxing both assumptions, we introduce a model whose likelihood can distinguish between rooted trees. The model is nonstationary with step changes in the instantaneous rate matrix at each speciation event. Exploiting recent theoretical work, each rate matrix belongs to a nonreversible family of Lie Markov models. These models are closed under matrix multiplication, so our extension offers the conceptually appealing property that a tree and all its subtrees could have arisen from the same family of nonstationary models.

We adopt a Bayesian approach, describe an MCMC algorithm for posterior inference and provide software. The biological insight that our model can provide is illustrated through an analysis in which nonreversible but stationary and nonstationary but reversible models cannot identify a plausible root.

REFERENCES

- BARRY, D. and HARTIGAN, J. A. (1987). Statistical analysis of hominoid molecular evolution. *Statist. Sci.* **2** 191–210. [MR0904034](https://doi.org/10.2307/2346134)
- BLANQUART, S. and LARTILLOT, N. (2006). A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* **23** 2058–2071. <https://doi.org/10.1093/molbev/msl091>
- BOHLIN, J., ELDHOLM, V., PETTERSSON, J. H. O., BRYNILDSRUD, O. and SNIPEN, L. (2017). The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics* **18** 151. <https://doi.org/10.1186/s12864-017-3543-7>
- BRAUN, E. L. and KIMBALL, R. T. (2002). Examining basal avian divergences with mitochondrial sequences: Model complexity, taxon sampling, and sequence length. *Syst. Biol.* **51** 614–625. <https://doi.org/10.1080/10635150290102294>
- BRYANT, D. (2003). A classification of consensus methods for phylogenetics. In *Bioconsensus* (Piscataway, NJ, 2000/2001). *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.* **61** 163–183. Amer. Math. Soc., Providence, RI. [MR1982426](https://doi.org/10.1090/PSI/003)
- CHERLIN, S., HEAPS, S., NYE, T., BOYS, R., WILLIAMS, T. and EMBLEY, T. (2018). The effect of non-reversibility on inferring rooted phylogenies. *Mol. Biol. Evol.* **35** 984–1002. <https://doi.org/10.1093/molbev/msx294>
- COX, C. J., FOSTER, P. G., HIRT, R. P., HARRIS, S. R. and EMBLEY, T. M. (2008). The archaeobacterial origin of eukaryotes. *Proc. Natl. Acad. Sci. USA* **105** 20356–20361.
- DUTHEIL, J. and BOUSSAU, B. (2008). Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol. Biol.* **28** 255.

- FELSENSTEIN, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* **22** 240–249.
- FERNÁNDEZ-SÁNCHEZ, J., SUMNER, J. G., JARVIS, P. D. and WOODHAMS, M. D. (2015). Lie Markov models with purine/pyrimidine symmetry. *J. Math. Biol.* **70** 855–891. [MR3306619 https://doi.org/10.1007/s00285-014-0773-z](https://doi.org/10.1007/s00285-014-0773-z)
- FOSTER, P. G. (2004). Modeling compositional heterogeneity. *Syst. Biol.* **53** 485–495. <https://doi.org/10.1080/10635150490445779>
- HANNAFORD, N. E., HEAPS, S. E., NYE, T. M., WILLIAMS, T. A. and EMBLEY, T. M. (2020). Supplement to “Incorporating compositional heterogeneity into Lie Markov models for phylogenetic inference.” <https://doi.org/10.1214/20-AOAS1369SUPPA>, <https://doi.org/10.1214/20-AOAS1369SUPPB>
- HASEGAWA, M., KISHINO, H. and YANO, T.-A. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22** 160–174. <https://doi.org/10.1007/BF02101694>
- HEAPS, S. E., NYE, T. M. W., BOYS, R. J., WILLIAMS, T. A. and EMBLEY, T. M. (2014). Bayesian modelling of compositional heterogeneity in molecular phylogenetics. *Stat. Appl. Genet. Mol. Biol.* **13** 589–609. [MR3276260 https://doi.org/10.1515/sagmb-2013-0077](https://doi.org/10.1515/sagmb-2013-0077)
- HODGKINSON, A. and EYRE-WALKER, A. (2011). Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12** 756–766.
- HUELSENBECK, J. P., BOLLBACK, J. P. and LEVINE, A. M. (2002). Inferring the root of a phylogenetic tree. *Syst. Biol.* **51** 32–43. <https://doi.org/10.1080/106351502753475862>
- KAehler, B. D. (2017). Full reconstruction of non-stationary strand-symmetric models on rooted phylogenies. *J. Theoret. Biol.* **420** 144–151. [MR3635328 https://doi.org/10.1016/j.jtbi.2017.03.007](https://doi.org/10.1016/j.jtbi.2017.03.007)
- LARTILLOT, N., BLANQUART, S. and LEPAGE, T. (2004). PhyloBayes 3.3: A Bayesian software for phylogenetic reconstruction and molecular dating using mixture models.
- LI, H., SHAO, R., SONG, N., SONG, F., JIANG, P., LI, Z. and CAI, W. (2015). Higher-level phylogeny of paraneopteran insects inferred from mitochondrial genome sequences. *Sci. Rep.* **5** 8527.
- LIND, P. A. and ANDERSSON, D. I. (2008). Whole-genome mutational biases in bacteria. *Proc. Natl. Acad. Sci. USA* **105** 17878–17883.
- MORGAN, C. C., FOSTER, P. G., WEBB, A. E., PISANI, D. and O’CONNELL, J. O. M. M. J. (2013). Heterogeneous models place the root of the placental mammal phylogeny. *Mol. Biol. Evol.* **30** 2145–2156.
- NEWTON, M. A. and RAFTERY, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. Roy. Statist. Soc. Ser. B* **56** 3–48. [MR1257793](https://doi.org/10.1093/sysbio/syz003)
- OAKS, J. R., COBB, K. A., MININ, V. N. and LEACHÉ, A. D. (2019). Marginal likelihoods in phylogenetics: A review of methods and applications. *Syst. Biol.* **68** 681–697. <https://doi.org/10.1093/sysbio/syz003>
- PHILLIPS, M. J., MCLENACHAN, P. A., DOWN, C., GIBB, G. C. and PENNY, D. (2006). Combined mitochondrial and nuclear DNA sequences resolve the interrelations of the major Australasian marsupial radiations. *Syst. Biol.* **55** 122–137. <https://doi.org/10.1080/106351505000481614>
- ROSENBERG, M. S., SUBRAMANIAN, S. and KUMAR, S. (2003). Patterns of transitional mutation biases within and among mammalian genomes. *Mol. Biol. Evol.* **20** 988–993.
- SQUARTINI, F. and ARNDT, P. F. (2008). Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Mol. Biol. Evol.* **25** 2525–2535.
- SUMNER, J. G., FERNÁNDEZ-SÁNCHEZ, J. and JARVIS, P. D. (2012). Lie Markov models. *J. Theoret. Biol.* **298** 16–31. [MR2899031 https://doi.org/10.1016/j.jtbi.2011.12.017](https://doi.org/10.1016/j.jtbi.2011.12.017)
- TARRÍO, R., RODRÍGUEZ-TRELLES, F. and AYALAA, F. J. (2000). Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: The *Drosophila* saltans and willistoni groups, a case study. *Mol. Phylogenet. Evol.* **16** 344–349.
- TAVARÉ, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In *Some Mathematical Questions in Biology—DNA Sequence Analysis* (New York, 1984). *Lectures Math. Life Sci.* **17** 57–86. Amer. Math. Soc., Providence, RI. [MR0846877](https://doi.org/10.1090/S0025-5718-1986-0084687-7)
- WILLIAMS, T. A., HEAPS, S. E., CHERLIN, S., NYE, T. M. W., BOYS, R. J. and EMBLEY, T. M. (2015). New substitution models for rooting phylogenetic trees. *Philos. Trans. R. Soc. B, Biol. Sci.* **370**.
- WOODHAMS, M. D., FERNÁNDEZ-SÁNCHEZ, J. and SUMNER, J. G. (2015). A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates. *Syst. Biol.* **64** 638–650.
- YANG, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39** 306–314. <https://doi.org/10.1007/BF00160154>
- YANG, Z. and ROBERTS, D. (1995). On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.* **12** 451–458. <https://doi.org/10.1093/oxfordjournals.molbev.a040220>
- ZWICKL, D. J. and HOLDER, M. T. (2004). Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. *Syst. Biol.* **53** 877–888.

HAWKES BINOMIAL TOPIC MODEL WITH APPLICATIONS TO COUPLED CONFLICT-TWITTER DATA

BY GEORGE MOHLER¹, ERIN MCGRATH², CODY BUNTAIN³ AND GARY LAFREE⁴

¹Department of Computer and Information Science, Indiana University—Purdue University Indianapolis, gmohler@iupui.edu

²National Consortium for the Study of Terrorism and Responses to Terrorism, University of Maryland,
McGrath.erinc@gmail.com

³Department of Informatics, New Jersey Institute of Technology, cbuntain@njit.edu

⁴Department of Criminology and Criminal Justice, University of Maryland, glafree@umd.edu

We consider the problem of modeling and clustering heterogeneous event data arising from coupled conflict event and social media data sets. In this setting conflict events trigger responses on social media, and, at the same time, signals of grievance detected in social media may serve as leading indicators for subsequent conflict events. For this purpose we introduce the Hawkes Binomial Topic Model (HBTM) where marks, Tweets and conflict event descriptions are represented as bags of words following a Binomial distribution. When viewed as a branching process, the daughter event bag of words is generated by randomly turning on/off parent words through independent Bernoulli random variables. We then use expectation–maximization to estimate the model parameters and branching structure of the process. The inferred branching structure is then used for topic cascade detection, short-term forecasting, and investigating the causal dependence of grievance on social media and conflict events in recent elections in Nigeria and Kenya.

REFERENCES

- AFROBAROMETER. Kenya and Nigeria, Round 6, 2014–2015.
- BELLIN, E. (2012). Reconsidering the robustness of authoritarianism in the Middle East: Lessons from the Arab Spring. *Comp. Polit.* **44** 127–149.
- BHAVNANI, R. and DONNAY, K. (2012). Here’s looking at you: The Arab Spring and violence in Gaza, Israel and the West Bank. *Swiss Pol. Sci. Rev.* **18** 124–131.
- BUNTAIN, C., GOLBECK, J., LIU, B. and LAFREE, G. (2016). Evaluating public response to the Boston marathon bombing and other acts of terrorism through Twitter. In *Tenth International AAAI Conference on Web and Social Media*.
- CHA, M., HADDADI, H., BENEVENUTO, F. and GUMMADI, K. P. (2010). Measuring user influence in Twitter: The million follower fallacy. In *Fourth International AAAI Conference on Weblogs and Social Media*. Available at <http://aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1538/0>.
- CHEESEMAN, N., LYNCH, G. and WILLIS, J. (2014). Democracy and its discontents: Understanding Kenya’s 2013 elections. *J. East. Afr. Stud.* **8** 2–24.
- DU, N., FARAJTABAR, M., AHMED, A., SMOLA, A. J. and SONG, L. (2015). Dirichlet–Hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 219–228. ACM, New York.
- DUTTA, H. S., DUTTA, V. R., ADHIKARY, A. and CHAKRABORTY, T. (2020). Hawkeseye: Detecting fake retweeters using Hawkes process and topic modeling. *IEEE Trans. Inform. Forensics Secur.* **15** 2667–2678.
- ECK, K. (2012). In data we trust? A comparison of UCDP GED and ACLED conflict events datasets. *Coop. Confl.* **47** 124–141.
- FARAJTABAR, M., WANG, Y., GOMEZ-RODRIGUEZ, M., LI, S., ZHA, H. and SONG, L. (2017). COEVOLVE: A joint point process model for information diffusion and network evolution. *J. Mach. Learn. Res.* **18** Paper No. 41, 49. [MR3655306](https://arxiv.org/abs/1703.03024)
- FREES, E. W., MEYERS, G. and CUMMINGS, A. D. (2011). Summarizing insurance scores using a Gini index. *J. Amer. Statist. Assoc.* **106** 1085–1098. [MR2894766 https://doi.org/10.1198/jasa.2011.tm10506](https://doi.org/10.1198/jasa.2011.tm10506)

- GOUWS, S., METZLER, D., CAI, C., HOVY, E. and REY, M. (2011). Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop of Language in Social Media*.
- GURR, T. (1968). Psychological factors in civil violence. *World Polit.* **20** 245–278.
- HE, X., REKATSINAS, T., FOULDS, J., GETOOR, L. and HAWKESTOPIC, Y. L. (2015). A joint model for network inference and topic modeling from text-based cascades. In *Proc. 32nd Intl. Conf. on Machine Learning*.
- KIM, M., PAINI, D. and JURDAK, R. (2020). Real-world diffusion dynamics based on point process approaches: A review. *Artif. Intell. Rev.* **53** 321–350.
- KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* **22** 79–86. [MR0039968 https://doi.org/10.1214/aoms/1177729694](https://doi.org/10.1214/aoms/1177729694)
- KWAK, H., LEE, C., PARK, H. and MOON, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web* 591–600. ACM, New York.
- LAI, E. L., MOYER, D., YUAN, B., FOX, E., HUNTER, B., BERTOZZI, A. L. and BRANTINGHAM, P. J. (2016). Topic time series analysis of microblogs. *IMA J. Appl. Math.* **81** 409–431. [MR3564661 https://doi.org/10.1093/imaamat/hxw025](https://doi.org/10.1093/imaamat/hxw025)
- 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. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. [MR2816705 https://doi.org/10.1198/jasa.2011.ap09546](https://doi.org/10.1198/jasa.2011.ap09546)
- MOHLER, G., MCGRATH, E., BUNTAİN, C. and LAFREE, G. (2020). Supplement to “Hawkes binomial topic model with applications to coupled conflict-Twitter data.” <https://doi.org/10.1214/20-AOAS1352SUPP>
- NEWMAN, D., LAU, J. H., GRIESER, K. and BALDWIN, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* 100–108. Association for Computational Linguistics, Stroudsburg.
- OWEN, O. and USMAN, Z. (2015). Why goodluck jonathan lost the nigerian presidential election of 2015. *Afr. Aff.* adv037.
- PORTER, M. D. (2016). A statistical approach to crime linkage. *Amer. Statist.* **70** 152–165. [MR3511045 https://doi.org/10.1080/00031305.2015.1123185](https://doi.org/10.1080/00031305.2015.1123185)
- PORTER, M. D. and WHITE, G. (2012). Self-exciting hurdle models for terrorist activity. *Ann. Appl. Stat.* **6** 106–124. [MR2951531 https://doi.org/10.1214/11-AOAS513](https://doi.org/10.1214/11-AOAS513)
- RALEIGH, C., LINKE, A., HEGRE, H. and KARLSEN, J. (2010). Introducing ACLED-armed conflict location and event data. *J. Peace Res.* **47** 651–660.
- RENNICK, S. A. (2013). Personal grievance sharing, frame alignment, and hybrid organisational structures: The role of social media in North Africa’s 2011 uprisings. *J. Contemp. Afr. Stud.* **31** 156–174.
- SIMMA, A. and JORDAN, M. I. (2012). Modeling events with cascades of Poisson processes. Preprint. Available at [arXiv:1203.3516](https://arxiv.org/abs/1203.3516).
- TAN, X., RAO, V. and NEVILLE, J. (2018). The Indian Buffet Hawkes process to model evolving latent influences. In *UAI* 795–804.
- 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](https://doi.org/10.1198/016214508000000148)
- XIE, W., ZHU, F., JIANG, J., LIM, E.-P. and TOPICKETCH, K. W. (2016). Real-time bursty topic detection from Twitter. *IEEE Trans. Knowl. Data Eng.* **28** 2216–2229.
- XU, H. and ZHA, H. (2017). A Dirichlet mixture model of Hawkes processes for event sequence clustering. In *Advances in Neural Information Processing Systems* 1354–1363.
- ZAMMIT-MANGION, A., DEWAR, M., KADIRKAMANATHAN, V. and SANGUINETTI, G. (2012). Point process modelling of the Afghan war diary. *Proc. Natl. Acad. Sci. USA* **109** 12414–12419.
- ZAMMIT-MANGION, A., DEWAR, M., KADIRKAMANATHAN, V., FLESKEN, A. and SANGUINETTI, G. (2013). Conflict data sets and point patterns. In *Modeling Conflict Dynamics with Spatio-Temporal Data. Springer-Briefs in Applied Sciences and Technology* 1–14. Springer, Cham.
- ZEITZOFF, T. (2011). Using social media to measure conflict dynamics: An application to the 2008–2009 Gaza conflict. *J. Confl. Resolut.* 0022002711408014.
- ZHAO, Q., ERDOGDU, M. A., HE, H. Y., RAJARAMAN, A. and SEISMIC, J. L. (2015). A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1513–1522. ACM, New York.
- ZHENG, P., YUAN, S., WU, X. and WU, Y. (2019). Identifying hidden buyers in darknet markets via Dirichlet Hawkes process. Preprint. Available at [arXiv:1911.04620](https://arxiv.org/abs/1911.04620).
- 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](https://doi.org/10.1198/016214502760046925)

NONPARAMETRIC BAYESIAN MULTIARMED BANDITS FOR SINGLE-CELL EXPERIMENT DESIGN

BY FEDERICO CAMERLENGHI¹, BIANCA DUMITRASCU², FEDERICO FERRARI³,
BARBARA E. ENGELHARDT⁴ AND STEFANO FAVARO⁵

¹Department of Economics, Management and Statistics, University of Milano – Bicocca, federico.camerlenghi@unimib.it

²School of Mathematics, Institute for Advanced Study, biancad@ias.edu

³Department of Statistical Science, Duke University, ff31@duke.edu

⁴Department of Computer Science, Center for Statistics and Machine Learning, Princeton University, bee@princeton.edu

⁵Department of Economics and Statistics, University of Torino, stefano.favaro@unito.it

The problem of maximizing cell type discovery under budget constraints is a fundamental challenge for the collection and analysis of single-cell RNA-sequencing (scRNA-seq) data. In this paper we introduce a simple, computationally efficient and scalable Bayesian nonparametric sequential approach to optimize the budget allocation when designing a large-scale experiment for the collection of scRNA-seq data for the purpose of, but not limited to, creating cell atlases. Our approach relies on the following tools: (i) a hierarchical Pitman–Yor prior that recapitulates biological assumptions regarding cellular differentiation, and (ii) a Thompson sampling multiarmed bandit strategy that balances exploitation and exploration to prioritize experiments across a sequence of trials. Posterior inference is performed by using a sequential Monte Carlo approach which allows us to fully exploit the sequential nature of our species sampling problem. We empirically show that our approach outperforms state-of-the-art methods and achieves near-Oracle performance on simulated and scRNA-seq data alike. HPY-TS code is available at <https://github.com/fedfer/HPYsinglecell>.

REFERENCES

- ANGERER, P., SIMON, L., TRITSCHLER, S., WOLF, F. A., FISCHER, D. and THEIS, F. J. (2017). Single cells make big data: New challenges and opportunities in transcriptomics. *Current Opinion in Systems Biology* **4** 85–91.
- AUER, P., CESA-BIANCHI, N. and FISCHER, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47** 235–256.
- BASSETTI, F., CASARIN, R. and ROSSINI, L. (2020). Hierarchical species sampling models. *Bayesian Anal.* **15** 809–838. <https://doi.org/10.1214/19-BA1168>
- BATTISTON, M., FAVARO, S. and TEH, Y. W. (2018). Multiarmed bandit for species discovery: A Bayesian nonparametric approach. *J. Amer. Statist. Assoc.* **113** 455–466. [MR3803478 https://doi.org/10.1080/01621459.2016.1261711](https://doi.org/10.1080/01621459.2016.1261711)
- BUBECK, S., ERNST, D. and GARIVIER, A. (2013). Optimal discovery with probabilistic expert advice: Finite time analysis and macroscopic optimality. *J. Mach. Learn. Res.* **14** 601–623. [MR3033341](https://doi.org/10.1214/17-AOS1678)
- CAMERLENGHI, F., LIJOI, A., ORBANZ, P. and PRÜNSTER, I. (2019). Distribution theory for hierarchical processes. *Ann. Statist.* **47** 67–92. [MR3909927 https://doi.org/10.1214/17-AOS1678](https://doi.org/10.1214/17-AOS1678)
- CAMERLENGHI, F., DUMITRASCU, B., FERRARI, F., ENGELHARDT, B. E. and FAVARO, S. (2020a). Supplementary material to “Nonparametric Bayesian multiarmed bandits for single-cell experiment design.” <https://doi.org/10.1214/20-AOAS1370SUPPA>
- CAMERLENGHI, F., DUMITRASCU, B., FERRARI, F., ENGELHARDT, B. E. and FAVARO, S. (2020b). Source code for “Nonparametric Bayesian multiarmed bandits for single-cell experiment design.” <https://doi.org/10.1214/20-AOAS1370SUPPB>
- DIXIT, A., PARNAS, O., LI, B., CHEN, J., FULCO, C. P., JERBY-ARNON, L., MARJANOVIC, N. D., DIONNE, D., BURKS, T. et al. (2016). Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167** 1853–1866.

Key words and phrases. Cell type discovery, experimental sampling design, hierarchical Pitman–Yor model, multiarmed bandits, scRNA-seq, sequential Monte Carlo, Thompson sampling.

- DUMITRASCU, B., FENG, K. and ENGELHARDT, B. E. (2018a). GT-TS: Experimental design for maximizing cell type discovery in single-cell data. *BioRxiv* 386540.
- DUMITRASCU, B., FENG, K. and ENGELHARDT, B. (2018b). Pólya–Gamma Thompson sampling for logistic contextual bandits. In *International Conference on Machine Learning*.
- EFRON, B. and THISTED, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63** 435–447.
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. [MR0350949](#)
- GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264. [MR0061330](#) <https://doi.org/10.1093/biomet/40.3-4.237>
- GOOD, I. J. and TOULMIN, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** 45–63. [MR0077039](#) <https://doi.org/10.1093/biomet/43.1-2.45>
- HAN, X., WANG, R., ZHOU, Y., FEI, L., SUN, H., LAI, S., SAADATPOUR, A., ZHOU, Z., CHEN, H. et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172** 1091–1107.
- HEAUKULANI, C., KNOWLES, D. and GHAHRAMANI, Z. (2014). Beta diffusion trees. In *International Conference on Machine Learning* 1809–1817.
- HORLBECK, M. A., XU, A., WANG, M., BENNETT, N. K., PARK, C. Y., BOGDANOFF, D., ADAMSON, B., CHOW, E. D., KAMPMANN, M. et al. (2018). Mapping the genetic landscape of human cells. *Cell* **174** 953–967.
- ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. [MR1952729](#) <https://doi.org/10.1198/016214501750332758>
- LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** 4–22. [MR0776826](#) [https://doi.org/10.1016/0196-8858\(85\)90002-8](https://doi.org/10.1016/0196-8858(85)90002-8)
- LIU, J. and WEST, M. (2001). Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice*. *Stat. Eng. Inf. Sci.* 197–223. Springer, New York. [MR1847793](#)
- MACOSKO, E. Z., BASU, A., SATIJA, R., NEMESH, J., SHEKHAR, K., GOLDMAN, M., TIROSH, I., BIALAS, A. R., KAMITAKI, N. et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161** 1202–1214.
- ORLITSKY, A., SURESH, A. T. and WU, Y. (2016). Optimal prediction of the number of unseen species. *Proc. Natl. Acad. Sci. USA* **113** 13283–13288. [MR3582444](#) <https://doi.org/10.1073/pnas.1607774113>
- PITMAN, J. and YOR, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25** 855–900. [MR1434129](#) <https://doi.org/10.1214/aop/1024404422>
- REGEV, A., TEICHMANN, S. A., LANDER, E. S., AMIT, I., BENOIST, C., BIRNEY, E., BODENMILLER, B., CAMPBELL, P., CARNINCI, P. et al. (2017). Science forum: The human cell atlas. *eLife* **6** e27041.
- RIZVI, A. H., CAMARA, P. G., KANDROR, E. K., ROBERTS, T. J., SCHIEREN, I., MANIATIS, T. and RABADAN, R. (2017). Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nat. Biotechnol.* **35** 551.
- SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. [MR1309433](#)
- SHIFFMAN, M., STEPHENSON, W. T., SCHIEBINGER, G., HUGGINS, J., CAMPBELL, T., REGEV, A. and BRODERICK, T. (2018). Reconstructing probabilistic trees of cellular differentiation from single-cell RNA-seq data. *arXiv preprint arXiv:1811.11790*.
- TEH, Y. W. (2006). A hierarchical Bayesian language model based on Pitman–Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* 985–992. Association for Computational Linguistics.
- TEH, Y. W. and JORDAN, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics*. *Camb. Ser. Stat. Probab. Math.* **28** 158–207. Cambridge Univ. Press, Cambridge. [MR2730663](#)
- THOMPSON, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25** 285–294.
- WADDINGTON, C. H. et al. (1957). *The Strategy of the Genes. A Discussion of Some Aspects of Theoretical Biology. With an Appendix by H. Kacser*. George Allen & Unwin, London.
- WEST, M. (1993a). Approximating posterior distributions by mixtures. *J. Roy. Statist. Soc. Ser. B* **55** 409–422. [MR1224405](#)
- WEST, M. (1993b). Mixture models, Monte Carlo, Bayesian updating, and dynamic models. *Computing Science and Statistics* 325–325.
- ZHENG, G. X., TERRY, J. M., BELGRADER, P., RYVKIN, P., BENT, Z. W., WILSON, R., ZIRALDO, S. B., WHEELER, T. D., MCDERMOTT, G. P. et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8** 14049.
- ZHU, L., LEI, J., DEVLIN, B. and ROEDER, K. (2018). A unified statistical framework for single cell and bulk RNA sequencing data. *Ann. Appl. Stat.* **12** 609–632. [MR3773407](#) <https://doi.org/10.1214/17-AOAS1110>

STRUCTURED DISCREPANCY IN BAYESIAN MODEL CALIBRATION FOR CHEMCAM ON THE MARS CURIOSITY ROVER

BY K. SHAM BHAT^{1,*}, KARY MYERS^{1,†}, EARL LAWRENCE^{1,‡}, JAMES COLGAN² AND ELIZABETH JUDGE³

¹Statistical Sciences, Los Alamos National Laboratory, *bhat9999@lanl.gov; †kary@lanl.gov; ‡earl@lanl.gov

²Theoretical Division, Los Alamos National Laboratory, jcolgan@lanl.gov

³Chemical Diagnostics and Engineering, Los Alamos National Laboratory, bethjudge@lanl.gov

The Mars rover Curiosity carries an instrument called ChemCam to determine the composition of the soil and rocks via laser-induced breakdown spectroscopy (LIBS). Los Alamos National Laboratory has developed a simulation capability that can predict spectra from ChemCam, but there are major-scale differences between the prediction and observation. This presents a challenge when using Bayesian model calibration to determine the unknown physical parameters that describe the LIBS observations. We present an analysis of LIBS data to support ChemCam based on including a structured discrepancy model in a Bayesian model-calibration scheme. This is both a novel application and an illustration of the importance of setting scientifically informed and constrained discrepancy models within Bayesian model calibration.

REFERENCES

- BRYNJARSDÓTTIR, J. and O'HAGAN, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Probl.* **30** 114007, 24. [MR3274591 https://doi.org/10.1088/0266-5611/30/11/114007](https://doi.org/10.1088/0266-5611/30/11/114007)
- COLGAN, J., JUDGE, E., KILCREASE, D. and BAREFIELD, J. II (2014). Ab-initio modeling of an iron laser-induced plasma: Comparison between theoretical and experimental atomic emission spectra. *Spectrochim. Acta, Part B: Atom. Spectrosc.* **97** 65–73.
- COLGAN, J., JUDGE, E., JOHNS, H., KILCREASE, D., BAREFIELD, J. II, MCINROY, R., HAKEL, P., WIENS, R. and CLEGG, S. (2015). Theoretical modeling and analysis of the emission spectra of a ChemCam standard: Basalt BIR-1A. *Spectrochim. Acta, Part B: Atom. Spectrosc.* **110** 20–30.
- COWAN, R. D. (1981). *The Theory of Atomic Structure and Spectra*. Number 3. Univ. California Press, Berkeley.
- FONTES, C., ZHANG, H., ABDALLAH, J. JR., CLARK, R., KILCREASE, D., COLGAN, J., CUNNINGHAM, R., HAKEL, P., MAGEE, N. et al. (2015). The Los Alamos suite of relativistic atomic physics codes. *J. Phys., B At. Mol. Opt. Phys.* **48** 144014.
- GATTIKER, J., MYERS, K., WILLIAMS, B., HIGDON, D., CARZOLIO, M. and HOEGH, A. (2016). Gaussian process-based sensitivity analysis and Bayesian model calibration with GPMSA. In *Handbook of Uncertainty Quantification* 1–41.
- GENOVESE, C. R., NOLL, D. C. and EDDY, W. F. (1997). Estimating test-retest reliability in functional MR imaging. I: Statistical methodology. *Magn. Reson. Med.* **38** 497–507. <https://doi.org/10.1002/mrm.1910380319>
- GRAY, G. III (2000). Classic split-hopkinson pressure bar testing. In *ASM Handbook* **8** 462–476.
- 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)
- 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)
- JOSEPH, V. R. and MELKOTE, S. N. (2009). Statistical adjustments to engineering models. *J. Qual. Technol.* **41** 362–375.
- JUDGE, E. J., COLGAN, J., CAMPBELL, K., BAREFIELD, J. E. II, JOHNS, H. M., KILCREASE, D. P. and CLEGG, S. (2016). Theoretical and experimental investigation of matrix effects observed in emission spectra of binary mixtures of sodium and copper and magnesium and copper pressed powders. *Spectrochim. Acta, Part B: Atom. Spectrosc.* **122** 142–148.

- KENNEDY, M. C. and O'HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. MR1858398 <https://doi.org/10.1111/1467-9868.00294>
- LEE, H., KASHYAP, V. L., VAN DYK, D. A., CONNORS, A., DRAKE, J. J., IZEM, R., MENG, X.-L., MIN, S., PARK, T. et al. (2011). Accounting for calibration uncertainties in X-ray analysis: Effective areas in spectral fitting. *Astrophys. J.* **731** 126.
- MAGEE, N. H., ABDALLAH, J., COLGAN, J., HAKEL, P., KILCREASE, D. P., MAZEVET, S., SHERRILL, M., FONTES, C. J. and ZHANG, H. (2004). Los Alamos Opacities: Transition from LEDCOP to ATOMIC. *AIP Conf. Proc.* **730** 168–179.
- MARMIN, S. and FILIPPONE, M. (2018). Variational calibration of computer models. Preprint. Available at [arXiv:1810.12177](https://arxiv.org/abs/1810.12177).
- MAURICE, S., WIENS, R., SACCOCCIO, M., BARRACLOUGH, B., GASNAULT, O., FORNI, O., MANGOLD, N., BARATOUX, D., BENDER, S. et al. (2012). The ChemCam instrument suite on the Mars Science Laboratory (MSL) rover: Science objectives and mast unit description. *Space Sci. Rev.* **170** 95–166.
- MENG, X.-L. (2018). Conducting highly principled data science: A statistician's job and joy. *Statist. Probab. Lett.* **136** 51–57. MR3806837 <https://doi.org/10.1016/j.spl.2018.02.053>
- MICHEL, A. P. and CHAVE, A. D. (2007). Analysis of laser-induced breakdown spectroscopy spectra: The case for extreme value statistics. *Spectrochim. Acta, Part B: Atom. Spectrosc.* **62** 1370–1378.
- NASA (2019). Curiosity rover mission overview. Available at <https://mars.nasa.gov/msl/mission/overview/>.
- 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>
- RAY, K. G. and MCCREERY, R. L. (1997). Simplified calibration of instrument response function for Raman spectrometers based on luminescent intensity standards. *Appl. Spectrosc.* **51** 108–116.
- SELLERS, K. F., MIECZNIKOWSKI, J., VISWANATHAN, S., MINDEN, J. S. and EDDY, W. F. (2007). Lights, camera, action! Systematic variation in 2-D difference gel electrophoresis images. *Electrophoresis* **28** 3324–3332.
- SJUE, S., AHRENS, J., BISWAS, A., FRANCOM, D., LAWRENCE, E., LUSCHER, D. and WALTERS, D. J. (2020). Fast strength model characterization using Bayesian statistics. In *Proceedings of the 21st Biennial Conference of the APS Topical Group on Shock Compression of Condensed Matter (SHOCK19)*.
- STOVER, F. S. and BRILL, R. V. (1998). Statistical quality control applied to ion chromatography calibrations. *J. Chromatogr. A* **804** 37–43.
- VAN DYK, D. A., CONNORS, A., KASHYAP, V. L. and SIEMIGINOWSKA, A. (2001). Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *Astrophys. J.* **548** 224.
- WIENS, R. C., MAURICE, S., BARRACLOUGH, B., SACCOCCIO, M., BARKLEY, W. C., BELL, J. F., BENDER, S., BERNARDIN, J., BLANEY, D. et al. (2012). The ChemCam instrument suite on the Mars Science Laboratory (MSL) rover: Body unit and combined system tests. *Space Sci. Rev.* **170** 167–227.
- WIENS, R., MAURICE, S., LASUE, J., FORNI, O., ANDERSON, R., CLEGG, S., BENDER, S., BLANEY, D., BARRACLOUGH, B. et al. (2013). Pre-flight calibration and initial data processing for the ChemCam laser-induced breakdown spectroscopy instrument on the Mars Science Laboratory rover. *Spectrochim. Acta, Part B: Atom. Spectrosc.* **82** 1–27.

MODELLING THE SOUND PRODUCTION OF NARWHALS USING A POINT PROCESS FRAMEWORK WITH MEMORY EFFECTS

BY ALEKSANDER SØLTOFT-JENSEN^{1,*}, MADPETER HEIDE-JØRGENSEN² AND
SUSANNE DITLEVSEN^{1,†}

¹*Data Science Laboratory, Department of Mathematical Sciences, University of Copenhagen, *soltoft-jensen@math.ku.dk;
†susanne@math.ku.dk*

²*Department of Birds and Mammals, Greenland Institute of Natural Resources, mhj@ghs.dk*

Obtaining an adequate description of the behaviour of narwhals in a pristine environment is important to understand natural behaviour as well as providing the means to determine potential changes in behaviour directly or indirectly caused by human activity. Based on Acousonde™ data from five narwhals in Scoresby Sound, this paper aims at modelling buzzing and calling rates of East Greenland narwhals as functions of time, space and, possibly, autoregressive memory. Both buzzing and calling are sounds produced by narwhals. Buzzing is a way for the whale to navigate and locate prey using echolocation, while calling is associated with social communication between whales. Logistic regression models without and with autoregressive components are compared based on AIC and comparatively assessed using diagnostics from point process theory. Adding an autoregressive component appears to improve the models, and further improvements for the buzzing model are made with a non-GLM extension. Effects of extrinsic covariates and memory are presented and interpreted. Buzzing occurs at deeper depths, and initiations of buzzes are separated by refractory periods. A possible feeding area is identified. Calling occurs closer to the surface, and, while the probability of calling in general is lower than buzzing, it is more likely that calls are clustered together rather than spread randomly.

REFERENCES

- BLACKWELL, S., TERVO, O., CONRAD, A., SINDING, M.-H., HANSEN, R., DITLEVSEN, S. and HEIDE-JØRGENSEN, M. P. (2018). Spatial and temporal patterns of sound production in East Greenland narwhals. *PLoS ONE* **13** 06.
- DERUITER, S. L., LANGROCK, R., SKIRBUTAS, T., GOLDBOGEN, J. A., CALAMBOKIDIS, J., FRIEDLAENDER, A. S. and SOUTHALL, B. L. (2017). A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. *Ann. Appl. Stat.* **11** 362–392. [MR3634328 https://doi.org/10.1214/16-AOAS1008](https://doi.org/10.1214/16-AOAS1008)
- HEIDE-JØRGENSEN, M. P., DIETZ, R., LAIDRE, K. and RICHARD, P. (2002). Autumn movements, home ranges, and winter density of narwhals (*Monodon monoceros*) tagged in Tremblay Sound, Baffin Island. *Polar Biol.* **25** 05.
- HEIDE-JØRGENSEN, M. P., NIELSEN, N., HANSEN, R. and BLACKWELL, S. (2014). Stomach temperature of narwhals (*Monodon monoceros*) during feeding events. *Anim. Biotelem.* **2** 06.
- HEIDE-JØRGENSEN, M. P., NIELSEN, N., HANSEN, R., SCHMIDT, H. C., BLACKWELL, S. and JØRGENSEN, O. A. (2015). The predictable narwhal: Satellite tracking shows behavioural similarities between isolated subpopulations. *J. Zool.* **297** 09.
- JOHNSON, M., AGUILAR DE SOTO, N. and MADSEN, P. T. (2009). Studying the behaviour and sensory ecology of marine mammals using acoustic recording tags: A review. *Mar. Ecol. Prog. Ser.* **395** 55.
- KASS, R. E., EDEN, U. T. and BROWN, E. N. (2014). *Analysis of Neural Data. Springer Series in Statistics.* Springer, New York. [MR3244261 https://doi.org/10.1007/978-1-4614-9602-1](https://doi.org/10.1007/978-1-4614-9602-1)
- KOBLITZ, J., STILZ, P., RASMUSSEN, M. and LAIDRE, K. L. (2016). Highly directional sonar beam of narwhals (*Monodon monoceros*) measured with a vertical 16 hydrophone array. *PLoS ONE* **11** 11.

- LANGROCK, R., MARQUES, T. A., BAIRD, R. W. and THOMAS, L. (2014). Modeling the diving behavior of whales: A latent-variable approach with feedback and semi-Markovian components. *J. Agric. Biol. Environ. Stat.* **19** 82–100. MR3257903 <https://doi.org/10.1007/s13253-013-0158-6>
- LI, K. and DITLEVSEN, S. (2019). Neural decoding with visual attention using sequential Monte Carlo for leaky integrate-and-fire neurons. *PLoS ONE* **14** 05.
- LI, K., KOZYREV, V., KYLLINGSBÆK, S., TREUE, S., DITLEVSEN, S. and BUNDESEN, C. (2016). Neurons in primate visual cortex alternate between responses to multiple stimuli in their receptive field. *Front. Comput. Neurosci.* **10** 12.
- NGO, M. C., HEIDE-JØRGENSEN, M. P. and DITLEVSEN, S. (2019). Understanding narwhal diving behaviour using hidden Markov models with dependent state distributions and long range dependence. *PLoS Comput. Biol.* **15**.
- NOREN, D., HOLT, M., DUNKIN, R. and WILLIAMS, T. (2017). Echolocation is cheap for some mammals: Dolphins conserve oxygen while producing high-intensity clicks. *J. Exp. Mar. Biol. Ecol.* **495** 103.
- PATTERSON, T. A., PARTON, A., LANGROCK, R., BLACKWELL, P. G., THOMAS, L. and KING, R. (2017). Statistical modelling of individual animal movement: An overview of key methods and a discussion of practical challenges. *AStA Adv. Stat. Anal.* **101** 399–438. MR3712406 <https://doi.org/10.1007/s10182-017-0302-7>
- QUICK, N. J., ISOJUNNO, S., SADYKOVA, D., BOWERS, M., NOWACEK, D. P. and HIDDEN, A. J. R. (2017). Markov models reveal complexity in the diving behaviour of short-finned pilot whales. *Sci. Rep.* **7**.
- RASMUSSEN, M., KOBLITZ, J. and LAIDRE, K. L. (2015). Buzzes and high-frequency clicks recorded from narwhals (*Monodon monoceros*) at their wintering ground. *Aquat. Mamm.* **41** 256.
- TRUCCOLO, W., EDEN, U. T., FELLOWS, M., DONOGHUE, J. 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** 03.
- WANG, J., TSANG, W. W. and MARSAGLIA, G. (2003). Evaluating Kolmogorov's distribution. *J. Stat. Softw.* **08** 12.
- YUAN, Y., BACHL, F. E., LINDGREN, F., BORCHERS, D. L., ILLIAN, J. B., BUCKLAND, S. T., RUE, H. and GERRODETTE, T. (2017). Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *Ann. Appl. Stat.* **11** 2270–2297. MR3743297 <https://doi.org/10.1214/17-AOAS1078>
- R CORE TEAM (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at <https://www.R-project.org/>.

REGION-REFERENCED SPECTRAL POWER DYNAMICS OF EEG SIGNALS: A HIERARCHICAL MODELING APPROACH

BY QIAN LI^{1,*}, JOHN SHAMSHOIAN¹, DAMLA ŞENTÜRK¹, CATHERINE SUGAR¹,
SHAFALI JESTE², CHARLOTTE DiSTEFANO² AND DONATELLO TELESCA^{1,†}

¹Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles,
^{*}qianl@ucla.edu; [†]dtelesca@ucla.edu

²Department of Psychiatry and Department of Psychiatry and Biobehavioral Sciences, UCLA David Geffen School of
Medicine, University of California, Los Angeles

Functional brain imaging through electroencephalography (EEG) relies upon the analysis and interpretation of high-dimensional, spatially organized time series. We propose to represent time-localized frequency domain characterizations of EEG data as region-referenced functional data. This representation is coupled with a hierarchical regression modeling approach to multivariate functional observations. Within this familiar setting we discuss how several prior models relate to structural assumptions about multivariate covariance operators. An overarching modeling framework, based on infinite factorial decompositions, is finally proposed to balance flexibility and efficiency in estimation. The motivating application stems from a study of implicit auditory learning, in which typically developing (TD) children, and children with autism spectrum disorder (ASD) were exposed to a continuous speech stream. Using the proposed model, we examine differential band power dynamics as brain function is interrogated throughout the duration of a computer-controlled experiment. Our work offers a novel look at previous findings in psychiatry and provides further insights into the understanding of ASD. Our approach to inference is fully Bayesian and implemented in a highly optimized Rcpp package.

REFERENCES

- BALADANDAYUTHAPANI, V., MALLICK, B. K. and CARROLL, R. J. (2005). Spatially adaptive Bayesian penalized regression splines (P-splines). *J. Comput. Graph. Statist.* **14** 378–394. [MR2160820 https://doi.org/10.1198/106186005X47345](https://doi.org/10.1198/106186005X47345)
- BALADANDAYUTHAPANI, V., MALLICK, B. K., HONG, M. Y., LUPTON, J. R., TURNER, N. D. and CARROLL, R. J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics* **64** 64–73, 321–322. [MR2422820 https://doi.org/10.1111/j.1541-0420.2007.00846.x](https://doi.org/10.1111/j.1541-0420.2007.00846.x)
- BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. [MR2806429 https://doi.org/10.1093/biomet/asr013](https://doi.org/10.1093/biomet/asr013)
- BRILLINGER, D. R. (1981). *Time Series: Data Analysis and Theory*, 2nd ed. *Holden-Day Series in Time Series Analysis*. Holden-Day, Inc., Oakland, CA. [MR0595684](https://doi.org/10.1198/0007123810885684)
- BROYD, S. J., DEMANUELE, C., DEBENER, S., HELPS, S. K., JAMES, C. J. and SONUGA-BARKE, E. J. (2009). Default-mode brain dysfunction in mental disorders: A systematic review. *Neurosci. Biobehav. Rev.* **33** 279–296.
- CHEN, K. and LYNCH, B. (2017). Weak separability for two-way functional data: Concept and test. arXiv preprint [arXiv:1703.10210](https://arxiv.org/abs/1703.10210).
- CHIOU, J.-M., CHEN, Y.-T. and YANG, Y.-F. (2014). Multivariate functional principal component analysis: A normalization approach. *Statist. Sinica* **24** 1571–1596. [MR3308652](https://doi.org/10.1198/106186007X208768)
- CRAINICEANU, C. M., RUPPERT, D., CARROLL, R. J., JOSHI, A. and GOODNER, B. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic errors. *J. Comput. Graph. Statist.* **16** 265–288. [MR2370943 https://doi.org/10.1198/106186007X208768](https://doi.org/10.1198/106186007X208768)
- CRESSIE, N. and HUANG, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Amer. Statist. Assoc.* **94** 1330–1340. [MR1731494 https://doi.org/10.2307/2669946](https://doi.org/10.2307/2669946)

- DAWID, A. P. (1985). Calibration-based empirical probability. *Ann. Statist.* **13** 1251–1285. With discussion by Mark J. Schervish and a reply by the author. [MR0811493 https://doi.org/10.1214/aos/1176349736](https://doi.org/10.1214/aos/1176349736)
- EIGSTI, I.-M., DE MARCHENA, A. B., SCHUH, J. M. and KELLEY, E. (2011). Language acquisition in autism spectrum disorders: A developmental review. *Research in Autism Spectrum Disorders* **5** 681–691.
- FINK, A. and BENEDEK, M. (2014). EEG alpha power and creative ideation. *Neurosci. Biobehav. Rev.* **44** 111–123.
- FRANKS, A. M. and HOFF, P. (2019). Shared subspace models for multi-group covariance estimation. *J. Mach. Learn. Res.* **20** Paper No. 171, 37. [MR4048982](https://doi.org/10.1214/aos/1176349736)
- GELMAN, A., HWANG, J. and VEHTARI, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24** 997–1016. [MR3253850 https://doi.org/10.1007/s11222-013-9416-2](https://doi.org/10.1007/s11222-013-9416-2)
- GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. With comments and a rejoinder by the authors. [MR1422404](https://doi.org/10.1007/s11222-013-9416-2)
- GOU, Z., CHOUDHURY, N. and BENASICH, A. A. (2011). Resting frontal gamma power at 16, 24 and 36 months predicts individual differences in language and cognition at 4 and 5 years. *Behav. Brain Res.* **220** 263–270.
- GREVEN, S., CRAINICEANU, C., CAFFO, B. and REICH, D. (2010). Longitudinal functional principal component analysis. *Electron. J. Stat.* **4** 1022–1054. [MR2727452 https://doi.org/10.1214/10-EJS575](https://doi.org/10.1214/10-EJS575)
- GRUBER, T. and MÜLLER, M. M. (2002). Effects of picture repetition on induced gamma band responses, evoked potentials, and phase synchrony in the human EEG. *Cogn. Brain Res.* **13** 377–392.
- HAPP, C. and GREVEN, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *J. Amer. Statist. Assoc.* **113** 649–659. [MR3832216 https://doi.org/10.1080/01621459.2016.1273115](https://doi.org/10.1080/01621459.2016.1273115)
- JAMES, G. M., HASTIE, T. J. and SUGAR, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87** 587–602. [MR1789811 https://doi.org/10.1093/biomet/87.3.587](https://doi.org/10.1093/biomet/87.3.587)
- KUHL, P. K. (2004). Early language acquisition: Cracking the speech code. *Nat. Rev., Neurosci.* **5** 831–843.
- LANG, S. and BREZGER, A. (2004). Bayesian P-splines. *J. Comput. Graph. Statist.* **13** 183–212. [MR2044877 https://doi.org/10.1198/1061860043010](https://doi.org/10.1198/1061860043010)
- LI, Q., SHAMSHOIAN, J., ŞENTÜRK, D., SUGAR, C., JESTE, S., DiSTEFANO, C. and TELESKA, D. (2020a). Supplement to “Region-referenced spectral power dynamics of EEG signals: A hierarchical modeling approach.” <https://doi.org/10.1214/20-AOAS1374SUPPA>
- LI, Q., SHAMSHOIAN, J., ŞENTÜRK, D., SUGAR, C., JESTE, S., DiSTEFANO, C. and TELESKA, D. (2020b). Source code for: “Region-referenced spectral power dynamics of EEG signals: A hierarchical modeling approach” <https://doi.org/10.1214/20-AOAS1374SUPPB>
- LORD, C., RISI, S., LAMBRECHT, L., COOK, E. H., LEVENTHAL, B. L., DiLAFORE, P. C., PICKLES, A. and RUTTER, M. (2000). The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* **30** 205–223.
- LYNCH, B. and CHEN, K. (2018). A test of weak separability for multi-way functional data, with application to brain connectivity studies. *Biometrika* **105** 815–831. [MR3877867 https://doi.org/10.1093/biomet/asv048](https://doi.org/10.1093/biomet/asv048)
- MILLS, D. L., COFFEY-CORINA, S. A. and NEVILLE, H. J. (1993). Language acquisition and cerebral specialization in 20-month-old infants. *J. Cogn. Neurosci.* **5** 317–334.
- MONTAGNA, S., TOKDAR, S. T., NEELON, B. and DUNSON, D. B. (2012). Bayesian latent factor regression for functional and longitudinal data. *Biometrics* **68** 1064–1073. [MR3040013 https://doi.org/10.1111/j.1541-0420.2012.01788.x](https://doi.org/10.1111/j.1541-0420.2012.01788.x)
- MURIAS, M., WEBB, S. J., GREENSON, J. and DAWSON, G. (2007). Resting state cortical connectivity reflected in EEG coherence in individuals with autism. *Biological Psychiatry* **62** 270–273.
- NISHIMURA, A. and SUCHARD, M. A. (2018). Prior-preconditioned conjugate gradient method for accelerated Gibbs sampling in “large n & large p” sparse Bayesian regression. Technical Report, arXiv preprint [arXiv:1810.12437](https://arxiv.org/abs/1810.12437).
- OMBAO, H. and HO, M. R. (2006). Time-dependent frequency domain principal components analysis of multichannel non-stationary signals. *Comput. Statist. Data Anal.* **50** 2339–2360. [MR2225573 https://doi.org/10.1016/j.csda.2004.12.011](https://doi.org/10.1016/j.csda.2004.12.011)
- PARK, S. Y. and STAICU, A.-M. (2015). Longitudinal functional data analysis. *Stat* **4** 212–226. [MR3405402 https://doi.org/10.1002/sta4.89](https://doi.org/10.1002/sta4.89)
- ROJAS, D. C. and WILSON, L. B. (2014). γ -Band abnormalities as markers of autism spectrum disorders. *Biomark. Med.* **8** 353–368.
- SABY, J. N. and MARSHALL, P. J. (2012). The utility of EEG band power analysis in the study of infancy and early childhood. *Dev. Neuropsychol.* **37** 253–273.
- SCHUEFLER, A., TELESKA, D., LI, Q., SUGAR, C. A., DiSTEFANO, C., JESTE, S. and ŞENTÜRK, D. (2020). Hybrid principal components analysis for region-referenced longitudinal functional EEG data. *Biostatistics* **21** 139–157. [MR4043850 https://doi.org/10.1093/biostatistics/kxy034](https://doi.org/10.1093/biostatistics/kxy034)

- SHAMSHOIAN, J., SENTURK, D., JESTE, S. and TELESKA, D. (2019). Bayesian analysis of multidimensional functional data.
- UHLHAAS, P. J. and SINGER, W. (2010). Abnormal neural oscillations and synchrony in schizophrenia. *Nat. Rev., Neurosci.* **11** 100–113.
- VEHTARI, A., GELMAN, A. and GABRY, J. (2015). Pareto smoothed importance sampling. arXiv preprint [arXiv:1507.02646](https://arxiv.org/abs/1507.02646).
- VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27** 1413–1432. [MR3647105 https://doi.org/10.1007/s11222-016-9696-4](https://doi.org/10.1007/s11222-016-9696-4)
- WAKEFIELD, J. (2013). *Bayesian and Frequentist Regression Methods. Springer Series in Statistics*. Springer, New York. [MR3025225 https://doi.org/10.1007/978-1-4419-0925-1](https://doi.org/10.1007/978-1-4419-0925-1)
- WATANABE, S. (2013). A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14** 867–897. [MR3049492](https://doi.org/10.1007/978-1-4419-0925-1)
- YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. [MR2160561 https://doi.org/10.1198/016214504000001745](https://doi.org/10.1198/016214504000001745)
- YUAN, Y. and JOHNSON, V. E. (2012). Goodness-of-fit diagnostics for Bayesian hierarchical models. *Biometrics* **68** 156–164. [MR2909864 https://doi.org/10.1111/j.1541-0420.2011.01668.x](https://doi.org/10.1111/j.1541-0420.2011.01668.x)
- ZHOU, L., HUANG, J. Z., MARTINEZ, J. G., MAITY, A., BALADANDAYUTHAPANI, V. and CARROLL, R. J. (2010). Reduced rank mixed effects models for spatially correlated hierarchical functional data. *J. Amer. Statist. Assoc.* **105** 390–400. [MR2757206 https://doi.org/10.1198/jasa.2010.tm08737](https://doi.org/10.1198/jasa.2010.tm08737)
- ZHU, H., LI, R. and KONG, L. (2012). Multivariate varying coefficient model for functional responses. *Ann. Statist.* **40** 2634–2666. [MR3097615 https://doi.org/10.1214/12-AOS1045](https://doi.org/10.1214/12-AOS1045)

MITIGATING UNOBSERVED SPATIAL CONFOUNDING WHEN ESTIMATING THE EFFECT OF SUPERMARKET ACCESS ON CARDIOVASCULAR DISEASE DEATHS

BY PATRICK M. SCHNELL¹ AND GEORGIA PAPADOGEORGOU²

¹*Division of Biostatistics, College of Public Health, The Ohio State University, schnell.31@osu.edu*

²*Department of Statistics, University of Florida, gpapadogeorgou@ufl.edu*

Confounding by unmeasured spatial variables has received some attention in the spatial statistics and causal inference literatures, but concepts and approaches have remained largely separated. In this paper we aim to bridge these distinct strands of statistics by considering unmeasured spatial confounding within a causal inference framework and estimating effects using outcome regression tools popular within the spatial literature. First, we show how using spatially correlated random effects in the outcome model, an approach common among spatial statisticians, does not necessarily mitigate bias due to spatial confounding, a previously published but not universally known result. Motivated by the bias term of commonly-used estimators, we propose an affine estimator which addresses this deficiency. We discuss how unbiased estimation of causal parameters in the presence of unmeasured spatial confounding can only be achieved under an untestable set of assumptions which will often be application specific. We provide a set of assumptions which describe how the exposure and outcome of interest relate to the unmeasured variables, and we show that this set of assumptions is sufficient for identification of the causal effect based on the observed data when spatial dependencies can be represented by a ring graph. We implement our method using a fully Bayesian approach applicable to any type of outcome variable. This work is motivated by and used to estimate the effect of county-level limited access to supermarkets on the rate of cardiovascular disease deaths in the elderly across the whole continental United States. Even though standard approaches return null or protective effects, our approach uncovers evidence of unobserved spatial confounding and indicates that limited supermarket access has a harmful effect on cardiovascular mortality.

REFERENCES

- AMERICAN HEART ASSOCIATION (2019). Heart disease and stroke statistics—2019 update: A report from the American Heart Association. *Circulation* **139** e56–e528.
- ANTONELLI, J., PARMIGIANI, G. and DOMINICI, F. (2019). High-dimensional confounding adjustment using continuous spike and slab priors. *Bayesian Anal.* **14** 825–848. [MR3960772 https://doi.org/10.1214/18-BA1131](https://doi.org/10.1214/18-BA1131)
- BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, Boca Raton, FL. [MR3362184](https://doi.org/10.1201/9780203021831)
- BARON, R. M. and KENNY, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51** 1173–1182.
- BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. [MR3207983 https://doi.org/10.1093/restud/rdt044](https://doi.org/10.1093/restud/rdt044)
- BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. [MR0373208](https://doi.org/10.2307/2344738)
- CENTERS FOR DISEASE CONTROL (2018). Underlying Cause of Death 1999–2017 on CDC WONDER Online Database. <https://wonder.cdc.gov/ucd-icd10.html>.

Key words and phrases. Causal inference, cardiovascular disease, food access, Markov random field, spatial confounding, unmeasured confounding.

- CHUNG, Y., RABE-HESKETH, S. and CHOI, I.-H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Stat. Med.* **32** 4071–4089. [MR3102435](#) <https://doi.org/10.1002/sim.5821>
- CONGDON, P. (2013). Assessing the impact of socioeconomic variables on small area variations in suicide outcomes in England. *Int. J. Environ. Res. Public Health* **10** 158–177.
- D’AMOUR, A. (2019). On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)* **89**.
- DWYER-LINDGREN, L., MOKDAD, A. H., SREBOTNJAK, T., FLAXMAN, A. D., HANSEN, G. M. and MURRAY, C. J. (2014). Cigarette smoking prevalence in US counties: 1996–2012. *Popul. Health Metr.* **12** 5. <https://doi.org/10.1186/1478-7954-12-5>
- GELFAND, A. E., KOTTAS, A. and MACEACHERN, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100** 1021–1035. [MR2201028](#) <https://doi.org/10.1198/016214504000002078>
- GELFAND, A. E. and VOUNATSOU, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* **4** 11–15.
- 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>
- HAHN, P. R., CARVALHO, C. M., PUELZ, D. and HE, J. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal.* **13** 163–182. [MR3737947](#) <https://doi.org/10.1214/16-BA1044>
- HANKS, E. M., SCHLIEP, E. M., HOOTEN, M. B. and HOETING, J. A. (2015). Restricted spatial regression in practice: Geostatistical models, confounding, and robustness under model misspecification. *Environmetrics* **26** 243–254. [MR3340961](#) <https://doi.org/10.1002/env.2331>
- HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. Supplementary material available online. [MR2816546](#) <https://doi.org/10.1198/jcgs.2010.08162>
- HIRANO, K. and IMBENS, G. W. (2004). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. Wiley Ser. Probab. Stat. 73–84. Wiley, Chichester. [MR2134803](#) <https://doi.org/10.1002/0470090456.ch7>
- HODGES, J. S. and REICH, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *Amer. Statist.* **64** 325–334. [MR2758564](#) <https://doi.org/10.1198/tast.2010.10052>
- HUGHES, J. and HARAN, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 139–159. [MR3008275](#) <https://doi.org/10.1111/j.1467-9868.2012.01041.x>
- 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>
- KEELE, L., TITIUNIK, R. and ZUBIZARRETA, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *J. Roy. Statist. Soc. Ser. A* **178** 223–239. [MR3291769](#) <https://doi.org/10.1111/rssa.12056>
- KELLER, J. P. and SZPIRO, A. A. (2019). Selecting a scale for spatial confounding adjustment. arXiv preprint [arXiv:1909.11161](#).
- KELLI, H. M., HAMMADAH, M., AHMED, H., KO, Y.-A., TOPEL, M., SAMMAN-TAHHAN, A., AWAD, M., PATEL, K., MOHAMMED, K. et al. (2017). Association between living in food deserts and cardiovascular risk. *Circulation: Cardiovascular Quality and Outcomes* **10** e003532.
- KELLI, H. M., KIM, J. H., SAMMAN TAHHAN, A., LIU, C., KO, Y.-A., HAMMADAH, M., SULLIVAN, S., SANDESARA, P., ALKHODER, A. A. et al. (2019). Living in food deserts and adverse cardiovascular outcomes in patients with cardiovascular disease. *Journal of the American Heart Association* **8** e010694.
- KOTTAS, A., DUAN, J. A. and GELFAND, A. E. (2008). Modeling disease incidence data with spatial and spatio-temporal Dirichlet process mixtures. *Biom. J.* **50** 29–42. [MR2414636](#) <https://doi.org/10.1002/bimj.200610375>
- LARAIA, B. A., SIEGA-RIZ, A. M., KAUFMAN, J. S. and JONES, S. J. (2004). Proximity of supermarkets is positively associated with diet quality index for pregnancy. *Prev. Med.* **39** 869–875.
- LEE, D. and SARRAN, C. (2015). Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies. *Environmetrics* **26** 477–487. [MR3415567](#) <https://doi.org/10.1002/env.2348>
- MORLAND, K., WING, S., ROUX, A. D. and POOLE, C. (2002). Neighborhood characteristics associated with the location of food stores and food service places. *Am. J. Prev. Med.* **22** 23–29.
- OGBURN, E. L., SHPITSER, I. and TCHETGEN TCHETGEN, E. J. (2019). Comment on “Blessings of multiple causes” [MR4047282]. *J. Amer. Statist. Assoc.* **114** 1611–1615. [MR4047286](#) <https://doi.org/10.1080/01621459.2019.1689139>
- PACIOREK, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statist. Sci.* **25** 107–125. [MR2741817](#) <https://doi.org/10.1214/10-STS326>

- PAPADOGEORGOU, G., CHOIRAT, C. and ZIGLER, C. M. (2019). Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. *Biostatistics* **20** 256–272. MR3922132 <https://doi.org/10.1093/biostatistics/kxx074>
- PEARCE, J., HISCOCK, R., BLAKELY, T. and WITTEN, K. (2008). The contextual effects of neighbourhood access to supermarkets and convenience stores on individual fruit and vegetable consumption. *J. Epidemiol. Community Health* **62** 198–201.
- POWELL, L. M., AULD, M. C., CHALOUPKA, F. J., O'MALLEY, P. M. and JOHNSTON, L. D. (2007). Associations between access to food stores and adolescent body mass index. *Am. J. Prev. Med.* **33** S301–S307.
- ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect *Math. Model.* **7** 1393–1512. MR0877758 [https://doi.org/10.1016/0270-0255\(86\)90088-6](https://doi.org/10.1016/0270-0255(86)90088-6)
- ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90** 122–129. MR1325119
- ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1899138 <https://doi.org/10.1007/978-1-4757-3692-2>
- ROSENBAUM, P. R. and RUBIN, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Roy. Statist. Soc. Ser. B* **45** 212–218.
- ROSENBAUM, P. R. and RUBIN, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 <https://doi.org/10.1093/biomet/70.1.41>
- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.
- RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.
- RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. *Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720 <https://doi.org/10.1017/CBO9780511755453>
- SALLIS, J. F., NADER, P. R., RUPP, J. W., ATKINS, C. J. and WILSON, W. C. (1986). San Diego surveyed for heart-healthy foods and exercise facilities. In *Public Health Reports* (1974-) 216–219.
- SCHNELL, P. M. and PAPADOGEORGOU, G. (2020a). Supplement to “Mitigating unobserved confounding when estimating the effect of supermarket access on cardiovascular disease deaths.” <https://doi.org/10.1214/20-AOAS1377SUPPA>
- SCHNELL, P. M. and PAPADOGEORGOU, G. (2020b). Supplement data and code to “Mitigating unobserved confounding when estimating the effect of supermarket access on cardiovascular disease deaths.” <https://doi.org/10.1214/20-AOAS1377SUPPB>
- THADEN, H. and KNEIB, T. (2018). Structural equation models for dealing with spatial confounding. *Amer. Statist.* **72** 239–252. MR3836447 <https://doi.org/10.1080/00031305.2017.1305290>
- UNITED STATES DEPARTMENT OF AGRICULTURE ECONOMIC RESEARCH SERVICE (2012). Food Environment Atlas. <https://www.ers.usda.gov/data-products/food-environment-atlas>.
- VANDERWEELE, T. J. and DING, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Ann. Intern. Med.* **167** 268–274. <https://doi.org/10.7326/M16-2607>
- VERBITSKY-SAVITZ, N. and RAUDENBUSH, S. W. (2012). Causal inference under interference in spatial settings: A case study evaluating community policing program in Chicago. *Epidemiol. Methods* **1** 105–130.
- WILSON, A. and REICH, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics* **70** 852–861. MR3295746 <https://doi.org/10.1111/biom.12203>
- WON, J.-H., LIM, J., KIM, S.-J. and RAJARATNAM, B. (2013). Condition-number-regularized covariance estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 427–450. MR3065474 <https://doi.org/10.1111/j.1467-9868.2012.01049.x>
- ZHOU, T., ELLIOTT, M. R. and LITTLE, R. J. A. (2019). Penalized spline of propensity methods for treatment comparison. *J. Amer. Statist. Assoc.* **114** 1–19. MR3941229 <https://doi.org/10.1080/01621459.2018.1518234>
- ZIGLER, C. M. and DOMINICI, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *J. Amer. Statist. Assoc.* **109** 95–107. MR3180549 <https://doi.org/10.1080/01621459.2013.869498>

LETTER TO THE EDITOR: FITTING A FOLDED NORMAL DISTRIBUTION WITHOUT EM

BY IAIN L. MACDONALD

Centre for Actuarial Research, University of Cape Town, iaain.macdonald@uct.ac.za

The problem of fitting a folded normal distribution by maximum likelihood has been described as ‘not straightforward’, and alternatives such as EM proposed. We suggest here that it is in fact straightforward to fit such a distribution by direct numerical maximization of the likelihood. We demonstrate this in an example. The relevant R code is included.

REFERENCES

- ELANDT, R. C. (1961). The folded normal distribution: Two methods of estimating parameters from moments. *Technometrics* **3** 551–562. [MR0130738](#) <https://doi.org/10.2307/1266561>
- JOHNSON, N. L. (1962). The folded normal distribution: Accuracy of estimation by maximum likelihood. *Technometrics* **4** 249–256. [MR0137221](#) <https://doi.org/10.2307/1266622>
- JUNG, S., FOSKEY, M. and MARRON, J. S. (2011). Principal arc analysis on direct product manifolds. *Ann. Appl. Stat.* **5** 578–603. [MR2810410](#) <https://doi.org/10.1214/10-AOAS370>
- LEONE, F. C., NELSON, L. S. and NOTTINGHAM, R. B. (1961). The folded normal distribution. *Technometrics* **3** 543–550. [MR0130737](#) <https://doi.org/10.2307/1266560>
- TSAGRIS, M. and STEWART, C. (2020). A folded model for compositional data analysis. *Aust. N. Z. J. Stat.* **62** 249–277.

RESPONSE TO ‘FITTING A FOLDED NORMAL DISTRIBUTION WITHOUT EM’

BY SUNGKYU JUNG¹, MARK FOSKEY² AND J. S. MARRON³

¹*Department of Statistics, Seoul National University, sungkyu@snu.ac.kr*

²*Accuray Incorporated, mark.foskey@gmail.com*

³*Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, marron@unc.edu*

REFERENCES

- ELTZNER, B., HUCKEMANN, S. and MARDIA, K. V. (2018). Torus principal component analysis with applications to RNA structure. *Ann. Appl. Stat.* **12** 1332–1359. [MR3834306](#) <https://doi.org/10.1214/17-AOAS1115>
- JUNG, S., DRYDEN, I. L. and MARRON, J. S. (2012). Analysis of principal nested spheres. *Biometrika* **99** 551–568. [MR2966769](#) <https://doi.org/10.1093/biomet/ass022>
- JUNG, S., FOSKEY, M. and MARRON, J. S. (2011). Principal arc analysis on direct product manifolds. *Ann. Appl. Stat.* **5** 578–603. [MR2810410](#) <https://doi.org/10.1214/10-AOAS370>
- JUNG, S., FOSKEY, M. and MARRON, J. S. (2020). Supplement to “Response to ‘Fitting a folded normal distribution without EM’.” <https://doi.org/10.1214/20-AOAS1411SUPP>.
- KIM, B., SCHULZ, J. and JUNG, S. (2020). Kurtosis test of modality for rotationally symmetric distributions on hyperspheres. *J. Multivariate Anal.* **178** 104603, 17. [MR4076342](#) <https://doi.org/10.1016/j.jmva.2020.104603>
- SCEALY, J. L. and WELSH, A. H. (2011). Regression for compositional data by using distributions defined on the hypersphere. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 351–375. [MR2815780](#) <https://doi.org/10.1111/j.1467-9868.2010.00766.x>
- TSAGRIS, M. and STEWART, C. (2020). A folded model for compositional data analysis. *Aust. N. Z. J. Stat.* **62** 249–277.

REJOINER: FITTING A FOLDED NORMAL DISTRIBUTION WITHOUT EM

BY IAIN L. MACDONALD

Centre for Actuarial Research, University of Cape Town, ian.macdonald@uct.ac.za

REFERENCES

- ELTZNER, B., HUCKEMANN, S. and MARDIA, K. V. (2018). Torus principal component analysis with applications to RNA structure. *Ann. Appl. Stat.* **12** 1332–1359. [MR3834306](#) <https://doi.org/10.1214/17-AOAS1115>
- JUNG, S., FOSKEY, M. and MARRON, J. S. (2011). Principal arc analysis on direct product manifolds. *Ann. Appl. Stat.* **5** 578–603. [MR2810410](#) <https://doi.org/10.1214/10-AOAS370>

The Annals of Applied Statistics

Next Issues

- Late 19th-century navigational uncertainties and their influence on sea surface temperature estimates CHENGUANG DAI, DUO CHAN, PETER HUYBERS AND NATESH PILLAI
- Discovering interactions using covariate informed random partition models
GARRITT L. PAGE, FERNANDO A. QUINTANA AND GARY L. ROSNER
- BayICE: A Bayesian hierarchical model for semi-reference-based deconvolution of bulk transcriptomic data AN-SHUN TAI,
GEORGE CHIEN-CHENG TSENG AND WEN-PING HSIEH
- Model free estimation of graphical model using gene expression data
JENNY YANG, YANG LIU, YUFENG LIU AND WEI SUN
- Accounting for smoking in forecasting mortality and life expectancy
YICHENG LI AND ADRIAN E. RAFTERY
- Integrative network learning for multi-modality biomarker data
SHANGHONG XIE, DONGLIN ZENG AND YUANJIA WANG
- Analyzing second order stochasticity of neural spiking under stimuli-bundle exposure
CHRISTOPHER D. GLYNN, SURYA T. TOKDAR, AZEEM ZAMAN, VALERIA C. CARUSO,
JEFFREY T. MOHL, SHAWN M. WILLETT AND JENNIFER M. GROH
- Point process modeling of drug overdoses with heterogeneous and missing data
XUEYING LIU, JEREMY CARTER, BRAD RAY AND GEORGE MOHLER
- Spike-and-slab lasso biclustering
GEMMA MORAN, VERONIKA ROCKOVA AND EDWARD GEORGE
- 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 THAWEEHAI, DAVID E. ARTERBURN,
KAREN J. COLEMAN AND SEBASTIEN HANEUSE
- An evaluation of statistical methods for aggregate patterns of replication failure
JACOB M. SCHAUER, KAITLYN G. FITZGERALD, SARAH PEKO-SPICER,
MENA WHALEN, RIRITA ZEJNULLAHI AND LARRY V. HEDGES
- An approximate best prediction approach to small area estimation for sheet and rill erosion under informative sampling EMILY BERG AND JAE-KWANG KIM
- Feature selection for data integration with mixed multi-view data
YULIA BAKER, TIFFANY M. TANG AND GENEVERA I. ALLEN
- Random-effects meta-analysis of Phase I dose-finding studies using stochastic process priors
MORENO URSINO, CHRISTIAN RÖVER, SARAH ZOHAR AND TIM FRIEDE
- Hierarchical multidimensional scaling for the comparison of musical performance styles
ANNA K. YANCHENKO AND PETER D. HOFF
- 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
- 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 ELLENBERG, SEAN HENNESSY, CUI TAO AND YONG CHEN

Continued

The Annals of Applied Statistics

Next Issues—Continued

- 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
- The role of scale in the estimation of cell-type proportions
GREGORY J. HUNT AND JOHANN A. GAGNON-BARTSCH
- A regression discontinuity design for ordinal running variable: Evaluating Central Bank purchases of corporate bonds
FAN LI, ANDREA MERCATANTI, TANELI MAKINEN AND ANDREA SILVESTRINI
- A multiple imputation procedure for record linkage and causal inference to estimate the effects of home-delivered meals MINGYANG SHAN, KALI THOMAS AND ROEE GUTMAN
- High-fidelity hurricane surge forecasting using emulation and sequential experiments
MATTHEW PLUMLEE, TAYLOR G. ASHER, WON CHANG AND MATTHEW V. BILSKIE
- 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
- Estimation of cell lineage trees by maximum likelihood phylogenetics
JEAN FENG, WILLIAM DEWITT, AARON MCKENNA, NOAH SIMON, AMY WILLIS AND FREDERICK MATSEN
- 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
- An empirical Bayes change point model for transcriptome time course data
TIAN TIAN, RUIHUA CHENG AND ZHI WEI
- 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
- Model-based feature selection and clustering of RNA-seq data for unsupervised subtype discovery DAVID K. LIM, NAIM U. RASHID AND JOSEPH G. IBRAHIM
- 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

Continued

The Annals of Applied Statistics

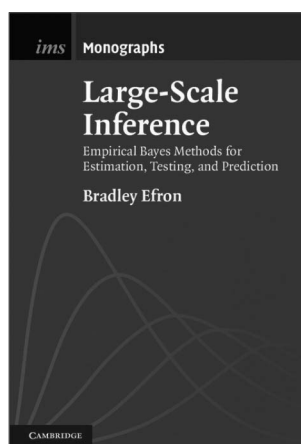
Next Issues—Continued

- 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



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.