

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

- Functional emulation of high resolution tsunami modelling over Cascadia
SERGE GUILLAS, ANDRIA SARRI, SIMON J. DAY,
XIAOYU LIU AND FREDERIC DIAS 2023
- Multi-rubric models for ordinal spatial data with application to online ratings
data ANTONIO R. LINERO, JONATHAN R. BRADLEY AND APURVA DESAI 2054
- The effects of nonignorable missing data on label-free mass spectrometry proteomics
experiments JONATHON J. O'BRIEN, HARSHA P. GUNAWARDENA,
JOAO A. PAULO, XIAN CHEN, JOSEPH G. IBRAHIM,
STEVEN P. GYGI AND BAHJAT F. QAQISH 2075
- Bottom-up estimation and top-down prediction: Solar energy prediction combining
information from multiple sources
YOUNGDEOK HWANG, SIYUAN LU AND JAE-KWANG KIM 2096
- ePCA: High dimensional exponential family PCA
LYDIA T. LIU, EDGAR DOBRIBAN AND AMIT SINGER 2121
- Rank tests in unmatched clustered randomized trials applied to a study of teacher
training PENG DING AND LUKE KEELE 2151
- Marked self-exciting point process modelling of information diffusion on Twitter
FENG CHEN AND WAI HONG TAN 2175
- Standardization of multivariate Gaussian mixture models and background adjustment of
PET images in brain oncology MENG LI AND ARMIN SCHWARTZMAN 2197
- Gaussian process modeling in approximate Bayesian computation to estimate horizontal
gene transfer in bacteria MARKO JÄRVENPÄÄ, MICHAEL U. GUTMANN,
AKI VEHTARI AND PEKKA MARTTINEN 2228
- A simulation-based framework for assessing the feasibility of respondent-driven sampling
for estimating characteristics in populations of lesbian, gay and bisexual older
adults MARYCLARE GRIFFIN, KRISTA J. GILE, KAREN I.
FREDRICKSEN-GOLDSSEN, MARK S. HANDCOCK AND ELENA A. EROSHEVA 2252
- Single stage prediction with embedded topic modeling of online reviews for mobile app
management SHAWN MANKAD, SHENGLI HU AND ANANDASIVAM GOPAL 2279
- Sensitivity analysis for stratified comparisons in an observational study of the effect of
smoking on homocysteine levels PAUL R. ROSENBAUM 2312
- Variable selection for estimating the optimal treatment regimes in the presence of a large
number of covariates BAQUN ZHANG AND MIN ZHANG 2335
- Modeling hybrid traits for comorbidity and genetic studies of alcohol and nicotine
co-dependence HEPING ZHANG, DUNGANG LIU, JIWEI ZHAO AND XUAN BI 2359
- Tree ensembles with rule structured horseshoe regularization
MALTE NALENZ AND MATTIAS VILLANI 2379
- Model transfer across additive manufacturing processes via mean effect equivalence of
lurking variables ARMAN SABBAGHI AND QIANG HUANG 2409

continued

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—Continued from front cover

- SCALPEL: Extracting neurons from calcium imaging data
ASHLEY PETERSEN, NOAH SIMON AND DANIELA WITTEN 2430
- Exact spike train inference via ℓ_0 optimization . . SEAN JEWELL AND DANIELA WITTEN 2457
- How often does the best team win? A unified approach to understanding randomness in
North American sport
MICHAEL J. LOPEZ, GREGORY J. MATTHEWS AND BENJAMIN S. BAUMER 2483
- Extending Bayesian structural time-series estimates of causal impact to many household
conservation initiatives
ERIC SCHMITT, CHRISTOPHER TULL AND PATRICK ATWATER 2517
- A locally adaptive process-convolution model for estimating the health impact of air
pollution DUNCAN LEE 2540
- Uncertainty through the lenses of a mixed-frequency Bayesian panel Markov-switching
model ROBERTO CASARIN, CLAUDIA FORONI, MASSIMILIANO MARCELLINO
AND FRANCESCO RAVAZZOLO 2559
- Disentangling and assessing uncertainties in multiperiod corporate default risk
predictions MIAO YUAN, CHENG YONG TANG, YILI HONG AND JIAN YANG 2587
- On the evolution of the United Kingdom price distributions
BA CHU, KIM HUYNH, DAVID JACHO-CHÁVEZ AND OLEKSIY KRYVTSOV 2618
- Instrumental variable analysis with censored data in the presence of many weak
instruments: Application to the effect of being sentenced to prison on time to
employment ASHKAN ERTEFAIE, ANH NGUYEN, DAVID J. HARDING,
JEFFREY D. MORENOFF AND WEI YANG 2647
- Variational inference for probabilistic Poisson PCA
JULIEN CHIQUET, MAHENDRA MARIADASSOU AND STÉPHANE ROBIN 2674
- Split-door criterion: Identification of causal effects through auxiliary outcomes
AMIT SHARMA, JAKE M. HOFMAN AND DUNCAN J. WATTS 2699

THE ANNALS OF APPLIED STATISTICS

Vol. 12, No. 4, pp. 2023–2733 December 2018

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: Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

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

Past President: Alison Etheridge, Department of Statistics, University of Oxford, Oxford, OX1 3LB, United Kingdom

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:* Edward I. George, Department of Statistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; Tailen Hsing, Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109-1107, USA

The Annals of Applied Statistics. *Editor-In-Chief:* Tilmann Gneiting, Heidelberg Institute for Theoretical Studies, HITS gGmbH, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany, and Institute for Stochastics, Karlsruhe Institute of Technology, Englerstr. 2, 76128 Karlsruhe, Germany

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. *Editor:* Bálint Tóth, School of Mathematics, University of Bristol, University Walk, BS8 1TW, Bristol, U.K. and Institute of Mathematics, Technical University Budapest, Egry József u. 1, H-1111 Budapest, Hungary

Statistical Science. *Editor:* Cun-Hui Zhang, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, USA

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 12, Number 4, December 2018. Published quarterly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, Ohio 44122, 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.

FUNCTIONAL EMULATION OF HIGH RESOLUTION TSUNAMI MODELLING OVER CASCADIA

BY SERGE GUILLAS^{*,1}, ANDRIA SARRI^{*}, SIMON J. DAY^{*,1}, XIAOYU LIU^{*}
 AND FREDERIC DIAS[†]

University College London^{} and University College Dublin[†]*

The rarity of tsunamis impels the scientific community to rely on numerical simulation for planning and risk assessment purposes because of the low availability of actual data from historic events. Numerical models, also called simulators, typically produce time series of outputs. Due to the large computational cost of such simulators, statistical emulation is required to carry out uncertainty quantification tasks, as emulators efficiently approximate simulators. There is thus a need to create emulators that respect the nature of time series outputs. We introduce here a novel statistical emulation of the input-output dependence of these computer models. We employ the Outer Product Emulator with two enhancements. Functional registration and Functional Principal Components techniques improve the predictions of the emulator. Our phase registration method captures fine variations in amplitude. Smoothness in the time series of outputs is modelled, and we are thus able to select more representative, and more parsimonious, regression functions than a fixed basis method such as a Fourier basis. We apply this approach to the high resolution tsunami wave propagation and coastal inundation for the Cascadia region in the Pacific Northwest. The coseismic representation in this analysis is novel, and more realistic than in previous studies. With the help of the emulator, we can carry out sensitivity analysis of the maximum wave elevation with respect to the source characteristics, and we are able to propagate uncertainties from the source characteristics to wave heights in order to issue probabilistic statements about tsunami hazard for Cascadia.

REFERENCES

- ADAMS, J. (1990). Paleoseismicity of the Cascadia subduction zone: Evidence from turbidites off the Oregon–Washington margin. *Tectonics* **9** 569–583.
- ATWATER, F. B., NELSON, R. A., CLAGUE, J. J., CARVER, A. G., YAMAGUCHI, K. D., BOBROWSKY, T. P., BOURGEOIS, J., DARIENZO, E. M., GRANT, C. W., HEMPHILL-HALEY, E., KELSEY, M. H., JACOBY, C. G., NISHENKO, P. S., PALMER, P. S., PETERSON, D. C. and REINHART, M. A. (1995). Summary of coastal geologic evidence for past great earthquakes at the Cascadia subduction zone. *Earthq. Spectra* **11** 1–18.
- BAYARRI, M. J., BERGER, J. O., CAPEO, J., GARCIA-DONATO, G., LIU, F., PALOMO, J., PARTHASARATHY, R. J., PAULO, R., SACKS, J. and WALSH, D. (2007). Computer model validation with functional output. *Ann. Statist.* **35** 1874–1906. [MR2363956](#)

Key words and phrases. Uncertainty quantification, emulation, tsunami modelling, sensitivity analysis, functional data analysis.

- BECK, J. and GUILLAS, S. (2016). Sequential design with mutual information for computer experiments (MICE): Emulation of a tsunami model. *SIAM/ASA J. Uncertain. Quantificat.* **4** 739–766. [MR3507556](#)
- BEHRENS, J. and DIAS, F. (2015). New computational methods in tsunami science. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **373** 20140382. [MR3405409](#)
- BERNARD, E. and TITOV, V. (2015). Evolution of tsunami warning systems and products. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **373** 20140371.
- BILIONIS, I., ZABARAS, N., KONOMI, B. A. and LIN, G. (2013). Multi-output separable Gaussian process: Towards an efficient, fully Bayesian paradigm for uncertainty quantification. *J. Comput. Phys.* **241** 212–239. [MR3647406](#)
- BRICKER, J. D. and NAKAYAMA, A. (2014). Contribution of trapped air, deck superelevation, and nearby structures to bridge deck failure during a tsunami. *J. Hydraul. Eng.* **140** 05014002.
- CASTRO, P. E., LAWTON, W. H. and SYLVESTRE, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28** 329–337.
- CHANG, W., APPEGATE, P. J., HARAN, M. and KELLER, K. (2014). Probabilistic calibration of a Greenland Ice Sheet model using spatially resolved synthetic observations: Toward projections of ice mass loss with uncertainties. *Geosci. Model Dev.* **7** 1933–1943.
- DAY, S. and FEARNLEY, C. (2015). A classification of mitigation strategies for natural hazards: Implications for the understanding of interactions between mitigation strategies. *Nat. Hazards* **79** 1219–1238.
- DUTYKH, D., PONCET, R. and DIAS, F. (2011). The VOLNA code for the numerical modeling of tsunami waves: Generation, propagation and inundation. *Eur. J. Mech. B Fluids* **30** 598–615. [MR2906226](#)
- GILES, M. B., MUDALIGE, G. R., SHARIF, Z., MARKALL, G. and KELLY, P. H. (2011). Performance analysis of the OP2 framework on many-core architectures. *ACM SIGMETRICS Perform. Eval. Rev.* **38** 9–15.
- GOLDFINGER, C., NELSON, C. H., JOHNSON, J. E., MOREY, A. E., GUTIERREZ-PASTOR, J., KARABANOV, E., ERIKSSON, A. T., GRACIA, E., DUNHILL, G., PATTON, J., ENKIN, R., DALLIMORE, A. and VALLIER, T. (2012). Turbidite event history: Methods and implications for Holocene paleoseismicity of the Cascadia Subduction Zone. U.S. Geol. Surv. Professional Paper 1661.
- GONZÁLEZ, F. I., GEIST, E. L., JAFFE, B., KÂNOĞLU, U., MOFJELD, H., SYNOLAKIS, C. E., TITOV, V. V., ARCAS, D., BELLOMO, D., CARLTON, D. et al. (2009). Probabilistic tsunami hazard assessment at seaside, Oregon, for near- and far-field seismic sources. *J. Geophys. Res., Oceans* **114** C11023.
- GUILLAS, S., SARRI, A., DAY, S. J., LIU, X. and DIAS, F. (2018). Supplement to “Functional emulation of high resolution tsunami modelling over Cascadia.” DOI:[10.1214/18-AOAS1142SUPP](#).
- GUSMAN, A. R., TANIOKA, Y., MACÍNNES, B. T. and TSUSHIMA, H. (2014). A methodology for near-field tsunami inundation forecasting: Application to the 2011 Tohoku tsunami. *J. Geophys. Res., Solid Earth* **119** 8186–8206.
- HAWKES, A. D., HORTON, B. P., NELSON, A. R., VANE, C. H. and SAWAI, Y. (2011). Coastal subsidence in Oregon, USA, during the giant Cascadia earthquake of AD 1700. *Quat. Sci. Rev.* **30** 364–376.
- HEIDARZADEH, M., PIROOZ, M. D. and ZAKER, N. H. (2009). Modeling the near-field effects of the worst-case tsunami in the Makran subduction zone. *Ocean Eng.* **36** 368–376.
- 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](#)
- HUNG, Y., JOSEPH, V. R. and MELKOTE, S. N. (2015). Analysis of computer experiments with functional response. *Technometrics* **57** 35–44. [MR3318347](#)
- HYNDMAN, R. D. (2013). Down dip landward limit of Cascadia great earthquake rupture. *J. Geophys. Res., Solid Earth* **118** 5530–5549.

- JOLLIFFE, I. T. (2002). *Principal Component Analysis*, 2nd ed. Springer, New York. [MR2036084](#)
- JOSEPH, V. R., HUNG, Y. and SUDJANTO, A. (2008). Blind kriging: A new method for developing metamodels. *J. Mech. Des.* **130** 031102.
- KAUFMAN, C. G., BINGHAM, D., HABIB, S., HEITMANN, K. and FRIEMAN, J. A. (2011). Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology. *Ann. Appl. Stat.* **5** 2470–2492. [MR2907123](#)
- KLEIBER, W., KATZ, R. W. and RAJAGOPALAN, B. (2013). Daily minimum and maximum temperature simulation over complex terrain. *Ann. Appl. Stat.* **7** 588–612. [MR3086432](#)
- KNEIP, A. and RAMSAY, J. O. (2008). Combining registration and fitting for functional models. *J. Amer. Statist. Assoc.* **103** 1155–1165. [MR2528838](#)
- LAY, T., KANAMORI, H., AMMON, C. J., NETTLES, M., WARD, S. N., ASTER, R. C., BECK, S. L., BILEK, S. L., BRUDZINSKI, M. R., BUTLER, R. et al. (2005). The great Sumatra-Andaman earthquake of 26 December 2004. *Science* **308** 1127–1133.
- LAY, T., AMMON, C. J., KANAMORI, H., KOPER, K. D., SUFRI, O. and HUTKO, A. R. (2010). Teleseismic inversion for rupture process of the 27 February 2010 Chile (Mw 8.8) earthquake. *Geophys. Res. Lett.* **37** L13301.
- LEONARD, L. J., ROGERS, G. C. and MAZZOTTI, S. (2014). Tsunami hazard assessment of Canada. *Nat. Hazards* **70** 237–274.
- LEONARD, L. J., CURRIE, C. A., MAZZOTTI, S. and HYNDMAN, R. D. (2010). Rupture area and displacement of past Cascadia great earthquakes from coastal coseismic subsidence. *Geol. Soc. Am. Bull.* **122** 2079–2096.
- LIU, X. and GUILLAS, S. (2017). Dimension reduction for Gaussian process emulation: An application to the influence of bathymetry on tsunami heights. *SIAM/ASA J. Uncertain. Quantificat.* **5** 787–812. [MR3686816](#)
- LÓPEZ-PINTADO, S. and ROMO, J. (2009). On the concept of depth for functional data. *J. Amer. Statist. Assoc.* **104** 718–734. [MR2541590](#)
- LUDWIN, R. S., DENNIS, R., CARVER, D., McMILLAN, A. D., LOSEY, R., CLAGUE, J., JONIENTZ-TRISLER, C., BOWECHOP, J., WRAY, J. and JAMES, K. (2005). Dating the 1700 Cascadia earthquake: Great coastal earthquakes in native stories. *Seismol. Res. Lett.* **76** 140–148.
- MASTERLARK, T. and HUGHES, K. L. (2008). Next generation of deformation models for the 2004 M9 Sumatra-Andaman earthquake. *Geophys. Res. Lett.* **35** L19310.
- MAZZOTTI, S., DRAGERT, H., HENTON, J., SCHMIDT, M., HYNDMAN, R., JAMES, T., LU, Y. and CRAYMER, M. (2003). Current tectonics of northern Cascadia from a decade of GPS measurements. *J. Geophys. Res., Solid Earth* **108** 2554.
- MOORE, G. F., BANGS, N. L., TAIRA, A., KURAMOTO, S., PANGBORN, E. and TOBIN, H. J. (2007). Three-dimensional splay fault geometry and implications for tsunami generation. *Science* **318** 1128–1131.
- MORRIS, M. D. (2012). Gaussian surrogates for computer models with time-varying inputs and outputs. *Technometrics* **54** 42–50. [MR2904734](#)
- MUDALIGE, G. R., GILES, M. B., REGULY, I., BERTOLLI, C. and KELLY, P. H. J. (2012). OP2: An active library framework for solving unstructured mesh-based applications on multi-core and many-core architectures. In 2012 *Innovative Parallel Computing (InPar)* 1–12. IEEE, Los Alamitos, CA.
- MUDALIGE, G. R., GILES, M. B., THIYAGALINGAM, J., REGULY, I. Z., BERTOLLI, C., KELLY, P. H. and TREFETHEN, A. E. (2013). Design and initial performance of a high-level unstructured mesh framework on heterogeneous parallel systems. *Parallel Comput.* **39** 669–692.
- NELSON, A. R., SAWAI, Y., JENNINGS, A. E., BRADLEY, L.-A., GERSON, L., SHERROD, B. L., SABEAN, J. and HORTON, B. P. (2008). Great-earthquake paleogeodesy and tsunamis of the past 2000 years at Alsea Bay, central Oregon coast, USA. *Quat. Sci. Rev.* **27** 747–768.
- OAKLEY, J. E. and O’HAGAN, A. (2004). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 751–769. [MR2088780](#)

- OKADA, Y. (1992). Internal deformation due to shear and tensile faults in a half-space. *Bull. Seismol. Soc. Am.* **82** 1018–1040.
- OKAL, E. A. (2015). The quest for wisdom: Lessons from 17 tsunamis, 2004–2014. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **373** 20140370.
- RAMSAY, J., HOOKER, G. and GRAVES, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer, Dordrecht.
- RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. [MR2168993](#)
- ROUGIER, J. (2008). Efficient emulators for multivariate deterministic functions. *J. Comput. Graph. Statist.* **17** 827–843. [MR2649069](#)
- ROUGIER, J., GUILLAS, S., MAUTE, A. and RICHMOND, A. D. (2009). Expert knowledge and multivariate emulation: The thermosphere-ionosphere electrodynamics general circulation model (TIE-GCM). *Technometrics* **51** 414–424. [MR2756477](#)
- SALMANIDOU, D. M., GUILLAS, S., GEORGIOPOULOU, A. and DIAS, F. (2017). Statistical emulation of landslide-induced tsunamis at the Rockall Bank, NE Atlantic. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **473** 20170026. [MR3650585](#)
- SARRI, A., GUILLAS, S. and DIAS, F. (2012). Statistical emulation of a tsunami model for sensitivity analysis and uncertainty quantification. *Nat. Hazards Earth Syst. Sci.* **12** 2003–2018.
- SATAKE, K., WANG, K. and ATWATER, B. F. (2003). Fault slip and seismic moment of the 1700 Cascadia earthquake inferred from Japanese tsunami descriptions. *J. Geophys. Res., Solid Earth* **108** 2535.
- SATAKE, K., SHIMAZAKI, K., TSUJI, Y. and UEDA, K. (1996). Time and size of a giant earthquake in Cascadia inferred from Japanese tsunami records of January 1700. *Nature* **379** 246–249.
- SATAKE, K., FUJII, Y., HARADA, T. and NAMEGAYA, Y. (2013). Time and space distribution of coseismic slip of the 2011 tohoku earthquake as inferred from tsunami waveform data. *Bull. Seismol. Soc. Am.* **103** 1473–1492.
- SILVERMAN, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Statist.* **24** 1–24. [MR1389877](#)
- SIMONS, M., MINSON, S. E., SLADEN, A. et al. (2011). The 2011 magnitude 9.0 Tohoku-Oki earthquake: Mosaicking the megathrust from seconds to centuries. *Science* **332** 1421–1425.
- SPILLER, E. T., BAYARRI, M. J., BERGER, J. O., CALDER, E. S., PATRA, A. K., PITMAN, E. B. and WOLPERT, R. L. (2014). Automating emulator construction for geophysical hazard maps. *SIAM/ASA J. Uncertain. Quantificat.* **2** 126–152. [MR3283903](#)
- WANG, K., MULDER, T., ROGERS, G. C. and HYNDMAN, R. D. (1995). Case for very low coupling stress on the Cascadia Subduction Fault. *J. Geophys. Res., Solid Earth* **100** 12907–12918.
- WANG, P.-L., ENGELHART, S. E., WANG, K., HAWKES, A. D., HORTON, B. P., NELSON, A. R. and WITTER, R. C. (2013). Heterogeneous rupture in the great Cascadia earthquake of 1700 inferred from coastal subsidence estimates. *J. Geophys. Res., Solid Earth* **118** 2460–2473.
- WELLS, D. L. and COPPERSMITH, K. J. (1994). New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bull. Seismol. Soc. Am.* **84** 974–1002.
- WITTER, R. C., ZHANG, Y. J., WANG, K., PRIEST, G. R., GOLDFINGER, C., STIMELY, L., ENGLISH, J. T. and FERRO, P. A. (2013). Simulated tsunami inundation for a range of Cascadia megathrust earthquake scenarios at Bandon, Oregon, USA. *Geosphere* **9** 1783–1803.

MULTI-RUBRIC MODELS FOR ORDINAL SPATIAL DATA WITH APPLICATION TO ONLINE RATINGS DATA

BY ANTONIO R. LINERO¹, JONATHAN R. BRADLEY² AND APURVA DESAI

Florida State University

Interest in online rating data has increased in recent years in which ordinal ratings of products or local businesses are provided by users of a website, such as Yelp! or Amazon. One source of heterogeneity in ratings is that users apply different standards when supplying their ratings; even if two users benefit from a product the same amount, they may translate their benefit into ratings in different ways. In this article we propose an ordinal data model, which we refer to as a multi-rubric model, which treats the criteria used to convert a latent utility into a rating as user-specific random effects, with the distribution of these random effects being modeled nonparametrically. We demonstrate that this approach is capable of accounting for this type of variability in addition to usual sources of heterogeneity due to item quality, user biases, interactions between items and users and the spatial structure of the users and items. We apply the model developed here to publicly available data from the website Yelp! and demonstrate that it produces interpretable clusterings of users according to their rating behavior, in addition to providing better predictions of ratings and better summaries of overall item quality.

REFERENCES

- ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. [MR1224394](#)
- ALBERT, J. H. and CHIB, S. (1997). Bayesian methods for cumulative, sequential and two-step ordinal data regression models. Technical report, Dept. Mathematics and Statistics, Bowling Green State Univ., Bowling Green, OH.
- BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 825–848. [MR2523906](#)
- BAO, J. and HANSON, T. E. (2015). Bayesian nonparametric multivariate ordinal regression. *Canad. J. Statist.* **43** 337–357. [MR3388321](#)
- BERRETT, C. and CALDER, C. A. (2012). Data augmentation strategies for the Bayesian spatial probit regression model. *Comput. Statist. Data Anal.* **56** 478–490. [MR2853748](#)
- BINDER, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65** 31–38. [MR0501592](#)
- BOBADILLA, J., ORTEGA, F., HERNANDO, A. and GUTIÉRREZ, A. (2013). Recommender systems survey. *Knowl.-Based Syst.* **46** 109–132.
- BRADLEY, J. R., HOLAN, S. H. and WIKLE, C. K. (2015). Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *Ann. Appl. Stat.* **9** 1761–1791. [MR3456353](#)
- BRADLEY, J. R., WIKLE, C. K. and HOLAN, S. H. (2016). Bayesian spatial change of support for count-valued survey data with application to the American community survey. *J. Amer. Statist. Assoc.* **111** 472–487. [MR3538680](#)

Key words and phrases. Bayesian hierarchical model, data augmentation, nonparametric Bayes, ordinal data, spatial prediction.

- CARGNONI, C., MÜLLER, P. and WEST, M. (1997). Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. *J. Amer. Statist. Assoc.* **92** 640–647.
- CARLIN, B. P. and POLSON, N. G. (1992). Monte Carlo Bayesian methods for discrete regression models and categorical time series. In *Bayesian Statistics, 4* (Peñíscola, 1991) (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. Smith, eds.) 109–125. Oxford Univ. Press, New York. [MR1380273](#)
- CHEN, M.-H. and DEY, D. K. (2000). A unified Bayesian approach for analyzing correlated ordinal response data. *Braz. J. Probab. Stat.* **14** 87–111. [MR1838453](#)
- COWLES, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Stat. Comput.* **6** 101–111.
- CRESSIE, N. A. C. (2015). *Statistics for Spatial Data*, revised ed. Wiley, New York. [MR3559472](#)
- CRESSIE, N. and JOHANNESSON, G. (2008). Fixed rank kriging for very large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 209–226. [MR2412639](#)
- DE OLIVEIRA, V. (2000). Bayesian prediction of clipped Gaussian random fields. *Comput. Statist. Data Anal.* **34** 299–314.
- DE OLIVEIRA, V. (2004). A simple model for spatial rainfall fields. *Stoch. Environ. Res. Risk Assess.* **18** 131–140.
- DEYOREO, M. and KOTTAS, A. (2014). Bayesian nonparametric modeling for multivariate ordinal regression. Preprint. Available at [arXiv:1408.1027](#).
- ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. [MR1340510](#)
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. [MR0350949](#)
- FRITSCH, A. and ICKSTADT, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Anal.* **4** 367–391. [MR2507368](#)
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. [MR2221284](#)
- GILL, J. and CASELLA, G. (2009). Nonparametric priors for ordinal Bayesian social science models: Specification and estimation. *J. Amer. Statist. Assoc.* **104** 453–464. [MR2751430](#)
- HÄRDLE, W. and MÜLLER, M. (2000). Multivariate and semiparametric kernel regression. In *Smoothing and Regression: Approaches, Computation, and Application* (M. G. Schimek, ed.). Wiley, New York.
- HIGGS, M. D. and HOETING, J. A. (2010). A clipped latent variable model for spatially correlated ordered categorical data. *Comput. Statist. Data Anal.* **54** 1999–2011. [MR2640303](#)
- HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.* **14** 1303–1347. [MR3081926](#)
- HOULSBY, N., HERNÁNDEZ-LOBATO, J. M. and GHAHRAMANI, Z. (2014). Cold-start active learning with robust ordinal matrix factorization. In *International Conference on Machine Learning* 766–774.
- ISHWARAN, H. and ZAREPOUR, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statist. Sinica* **12** 941–963. [MR1929973](#)
- KNORR-HELD, L. (1995). Dynamic cumulative probit models for ordinal panel-data; a Bayesian analysis by Gibbs sampling. Technical report, Ludwig-Maximilians-Univ., Munich.
- KOREN, Y. and SILL, J. (2011). OrdRec: An ordinal model for predicting personalized item rating distributions. In *Proceedings of the Fifth ACM Conference on Recommender Systems* 117–124. ACM, New York.
- KOTTAS, A., MÜLLER, P. and QUINTANA, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *J. Comput. Graph. Statist.* **14** 610–625. [MR2170204](#)
- LINERO, A. R. (2018). Bayesian regression trees for high dimensional prediction and variable selection. *J. Amer. Statist. Assoc.* **113** 626–636. [MR3832214](#)

- LINERO, A. R., BRADLEY, J. R. and DESAI, A. (2018). Supplement to “Multi-rubric models for ordinal spatial data with application to online ratings data”. DOI:[10.1214/18-AOAS1143SUPP](https://doi.org/10.1214/18-AOAS1143SUPP).
- MARLIN, B. M. and ZEMEL, R. S. (2009). Collaborative prediction and ranking with non-random missing data. In *Proceedings of the Third ACM Conference on Recommender Systems* 5–12. ACM, New York.
- MARLIN, B. M., ZEMEL, R. S., ROWEIS, S. and SLANEY, M. (2007). Collaborative filtering and the missing at random assumption. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence* 267–275. AUAI Press, Corvallis, OR.
- MNIH, A. and SALAKHUTDINOV, R. R. (2008). Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems* **20** 1257–1264. Curran Associates, Red Hook, NY.
- NEAL, R. M. (2003). Slice sampling. *Ann. Statist.* **31** 705–767. [MR1994729](#)
- PAQUET, U., THOMSON, B. and WINTHER, O. (2012). A hierarchical model for ordinal matrix factorization. *Stat. Comput.* **22** 945–957. [MR2913794](#)
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA. [MR2514435](#)
- SANG, H. and HUANG, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 111–132. [MR2885842](#)
- SCHLIEP, E. M. and HOETING, J. A. (2015). Data augmentation and parameter expansion for independent or spatially correlated ordinal data. *Comput. Statist. Data Anal.* **90** 1–14. [MR3354825](#)
- SIMPSON, D., RUE, H., RIEBLER, A., MARTINS, T. G. and SØRBYE, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.* **32** 1–28. [MR3634300](#)
- TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. [MR2279480](#)
- VELOZO, P. L., ALVES, M. B. and SCHMIDT, A. M. (2014). Modelling categorized levels of precipitation. *Braz. J. Probab. Stat.* **28** 190–208. [MR3189493](#)

THE EFFECTS OF NONIGNORABLE MISSING DATA ON LABEL-FREE MASS SPECTROMETRY PROTEOMICS EXPERIMENTS¹

BY JONATHON J. O'BRIEN*, HARSHA P. GUNAWARDENA[†], JOAO A. PAULO*,
XIAN CHEN[†], JOSEPH G. IBRAHIM[†], STEVEN P. GYGI* AND
BAHJAT F. QAQISH[†]

Harvard Medical School and University of North Carolina at Chapel Hill[†]*

An idealized version of a label-free discovery mass spectrometry proteomics experiment would provide absolute abundance measurements for a whole proteome, across varying conditions. Unfortunately, this ideal is not realized. Measurements are made on peptides requiring an inferential step to obtain protein level estimates. The inference is complicated by experimental factors that necessitate relative abundance estimation and result in widespread nonignorable missing data. Relative abundance on the log scale takes the form of parameter contrasts. In a complete-case analysis, contrast estimates may be biased by missing data, and a substantial amount of useful information will often go unused.

To avoid problems with missing data, many analysts have turned to single imputation solutions. Unfortunately, these methods often create further difficulties by hiding inestimable contrasts, preventing the recovery of interblock information and failing to account for imputation uncertainty. To mitigate many of the problems caused by missing values, we propose the use of a Bayesian selection model. Our model is tested on simulated data, real data with simulated missing values, and on a ground truth dilution experiment where all of the true relative changes are known. The analysis suggests that our model, compared with various imputation strategies and complete-case analyses, can increase accuracy and provide substantial improvements to interval coverage.

REFERENCES

- AZZALINI, A. (2014). *The Skew-Normal and Related Families*. Institute of Mathematical Statistics (IMS) Monographs **3**. Cambridge Univ. Press, Cambridge. [MR3468021](#)
- CATHERMAN, A. D., SKINNER, O. S. and KELLEHER, N. L. (2014). Top down proteomics: Facts and perspectives. *Biochem. Biophys. Res. Commun.* **445** 683–693.
- CHEN, E. I. and YATES, J. R. (2007). Cancer proteomics by quantitative shotgun proteomics. *Mol. Oncol.* **1** 144–159.
- CLOUGH, T., THAMINY, S., RAGG, S., AEBERSOLD, R. and VITEK, O. (2012). Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. *BMC Bioinform.* **13**(Suppl. 1) S6.

Key words and phrases. Data dependent analysis, estimable contrasts, selection model, Bayesian inference, imputation, interval coverage.

- COX, J. and MANN, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26** 1367–1372.
- COX, J., HEIN, M. Y., LUBER, C. A., PARON, I., NAGARAJ, N. and MANN, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13** 2513–2526.
- DE BREVERN, A. G., HAZOUT, S. and MALPERTUY, A. (2004). Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinform.* **5** Art. ID 114.
- FABRE, B., LAMBOUR, T., BOUYSSIÉ, D., MENNETEAU, T., MONSARRAT, B., BURLET-SCHILTZ, O. and BOUSQUET-DUBOUCH, M.-P. (2014). Comparison of label-free quantification methods for the determination of protein complexes subunits stoichiometry. *EuPA Open Proteomics* **4** 82–86.
- KARPIEVITCH, Y. V., DABNEY, A. R. and SMITH, R. D. (2012). Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinform.* **13**(Suppl. 1) S5.
- KARPIEVITCH, Y., STANLEY, J., TAVERNER, T., HUANG, J., ADKINS, J. N., ANSONG, C., HEFFRON, F., METZ, T. O., QIAN, W. J., YOON, H., SMITH, R. D. and DABNEY, A. R. (2009). A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* **25** 2028–2034.
- KEILHAUER, E. C., HEIN, M. Y. and MANN, M. (2015). Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Mol. Cell. Proteomics* **14** 120–135.
- LAZAR, C., GATTO, L., FERRO, M., BRULEY, C. and BURGER, T. (2016). Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *J. Proteome Res.* **15** 1116–1125.
- LIEBLER, D. C. and ZIMMERMAN, L. J. (2013). Targeted quantitation of proteins by mass spectrometry. *Biochemistry* **52** 3797–3806.
- LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York. [MR0890519](#)
- LUO, R., COLANGELO, C. M., SESSA, W. C. and ZHAO, H. (2009). Bayesian analysis of iTRAQ data with nonrandom missingness: Identification of differentially expressed proteins. *Statist. Biosci.* **1** 228–245.
- MUELLER, L. N., RINNER, O., SCHMIDT, A., LETARTE, S., BODENMILLER, B., BRUSNIAK, M.-Y., VITEK, O., AEBERSOLD, R. and MÜLLER, M. (2007). SuperHirn—A novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7** 3470–3480.
- O'BRIEN, J. J., O'CONNELL, J. D., PAULO, J. A., THAKURTA, S., ROSE, C. M., WEEKES, M. P., HUTTLIN, E. L. and GYGI, S. P. (2018a). Compositional proteomics: Effects of spatial constraints on protein quantification utilizing isobaric tags. *J. Proteome Res.* **17** 590–599.
- O'BRIEN, J. J., GUNAWARDENA, H. P., PAULO, J. A., CHEN, X., IBRAHIM, J. G., GYGI, S. P. and QAQISH, B. F. (2018b). Supplement to “The effects of non-ignorable missing data on label-free mass spectrometry proteomics experiments.” DOI:[10.1214/18-AOAS1144SUPP](#).
- OWEN, A. B. and PERRY, P. O. (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Ann. Appl. Stat.* **3** 564–594. [MR2750673](#)
- POLPITIYA, A. D., QIAN, W.-J., JAITLY, N., PETYUK, V. A., ADKINS, J. N., CAMP, D. G., ANDERSON, G. A. and SMITH, R. D. (2008). DAnTE: A statistical tool for quantitative analysis of -omics data. *Bioinformatics* **24** 1556–1558.
- ROSS, P. L., HUANG, Y. N., MARCHESE, J. N., WILLIAMSON, B., PARKER, K., HATTAN, S., KHAINOVSKI, N., PILLAI, S., DEY, S., DANIELS, S., PURKAYASTHA, S., JUHASZ, P., MARTIN, S., BARTLET-JONES, M., HE, F., JACOBSON, A. and PAPPIN, D. J. (2004). Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3** 1154–1169.

- RÖST, H. L., ROSENBERGER, G., NAVARRO, P., GILLET, L., MILADINOVIĆ, S. M., SCHUBERT, O. T., WOLSKI, W., COLLINS, B. C., MALMSTRÖM, J., MALMSTRÖM, L. and AEBERSOLD, R. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **32** 219–223.
- SACCO, F., HUMPHREY, S. J., COX, J., MISCHNIK, M., SCHULTE, A., KLABUNDE, T., SCHÄFER, M. and MANN, M. (2016). Glucose-regulated and drug-perturbed phosphoproteome reveals molecular mechanisms controlling insulin secretion. *Nat. Commun.* **7** Art. ID 13250.
- SANDIN, M., KROGH, M., HANSSON, K. and LEVANDER, F. (2011). Generic workflow for quality assessment of quantitative label-free LC-MS analysis. *Proteomics* **11** 1114–1124.
- SCHEFFÉ, H. (1999). *The Analysis of Variance*. Wiley, New York. Reprint of the 1959 original. [MR1673563](#)
- SCHLIEKELMAN, P. and LIU, S. (2014). Quantifying the effect of competition for detection between coeluting peptides on detection probabilities in mass-spectrometry-based proteomics. *J. Proteome Res.* **13** 348–361.
- TAVERNER, T., KARPIEVITCH, Y. V., POLPITIYA, A. D., BROWN, J. N., DABNEY, A. R., ANDERSON, G. A. and SMITH, R. D. (2012). DanteR: An extensible R-based tool for quantitative analysis of -omics data. *Bioinformatics* **28** 2404–2406.
- TEKWE, C. D., CARROLL, R. J. and DABNEY, A. R. (2012). Application of survival analysis methodology to the quantitative analysis of LC-MS proteomics data. *Bioinformatics* **28** 1998–2003.
- THOMPSON, A., SCHÄFER, J., KUHN, K., KIENLE, S., SCHWARZ, J., SCHMIDT, G., NEUMANN, T. and HAMON, C. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75** 1895–1904.
- TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., BROWN, P., HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D. and ALTMAN, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* **17** 520–525.
- WEBB-ROBERTSON, B. J. M., WIBERG, H. K., MATZKE, M. M., BROWN, J. N., WANG, J., MCDERMOTT, J. E., SMITH, R. D., RODLAND, K. D., METZ, T. O., POUNDS, J. G. and WATERS, K. M. (2015). Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J. Proteome Res.* **14** 1993–2001.

BOTTOM-UP ESTIMATION AND TOP-DOWN PREDICTION: SOLAR ENERGY PREDICTION COMBINING INFORMATION FROM MULTIPLE SOURCES

BY YOUNGDEOK HWANG^{*,1,2}, SIYUAN LU^{†,1} AND JAE-KWANG KIM^{‡,§,3}

Sungkyunkwan University^{}, IBM Thomas J. Watson Research Center[†],
Iowa State University[‡] and KAIST[§]*

Accurately forecasting solar power using the data from multiple sources is an important but challenging problem. Our goal is to combine two different physics model forecasting outputs with real measurements from an automated monitoring network so as to better predict solar power in a timely manner. To this end, we propose a new approach of analyzing large-scale multilevel models with great computational efficiency requiring minimum monitoring and intervention. This approach features a division of the large scale data set into smaller ones with manageable sizes, based on their physical locations, and fit a local model in each area. The local model estimates are then combined sequentially from the specified multilevel models using our novel bottom-up approach for parameter estimation. The prediction, on the other hand, is implemented in a top-down matter. The proposed method is applied to the solar energy prediction problem for the U.S. Department of Energy's SunShot Initiative.

REFERENCES

- BATES, D. M. and PINHEIRO, J. C. (1998). Computational methods for multilevel modelling. Univ. Wisconsin, Madison, WI.
- 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](#)
- CHATTERJEE, S., LAHIRI, P. and LI, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *Ann. Statist.* **36** 1221–1245. [MR2418655](#)
- CHEN, X. and XIE, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statist. Sinica* **24** 1655–1684. [MR3308656](#)
- CHU, Y., PEDRO, H. T. C., NONNENMACHER, L., INMAN, R. H., LIAO, Z. and COIMBRA, C. F. M. (2014). A smart image-based cloud detection system for intrahour solar irradiance forecasts. *J. Atmos. Ocean. Technol.* **31** 1995–2007.
- DENHOLM, P. and MARGOLIS, R. M. (2007). Evaluating the limits of solar photovoltaics (PV) in traditional electric power systems. *Energy Policy* **35** 2852–2861.
- DU, J. and TRACTON, M. S. (2001). Implementation of a real-time shortrange ensemble forecasting system at NCEP: An update. In *Ninth Conference on Mesoscale Processes*. American Meteorological Society, Fort Lauderdale, FL.

Key words and phrases. Big data prediction, large-scale monitoring data, multilevel model, physics models, solar power generation.

- ELA, E., MILLIGAN, M. and KIRBY, B. (2011). Operating reserves and variable generation. NREL/TP-5500-51978. Available at <http://www.nrel.gov/docs/fy11osti/51978.pdf>.
- GELMAN, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics* **48** 432–435. [MR2252307](#)
- GELMAN, A., VEHTARI, A., JYLÄNKI, P., ROBERT, C., CHOPIN, N. and CUNNINGHAM, J. P. (2014). Expectation propagation as a way of life. Preprint. Available at [arXiv:1412.4869](#).
- GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73** 43–56. [MR0836433](#)
- GRAMACY, R. B. (2016). laGP: Large-scale spatial modeling via local approximate Gaussian processes in R. *J. Stat. Softw.* **72** 1–46.
- GRAMACY, R. B. and APLEY, D. W. (2015). Local Gaussian process approximation for large computer experiments. *J. Comput. Graph. Statist.* **24** 561–578. [MR3357395](#)
- GRAMACY, R. B. and HAALAND, B. (2016). Speeding up neighborhood search in local Gaussian process prediction. *Technometrics* **58** 294–303. [MR3520659](#)
- GRAMACY, R. B., BINGHAM, D., HOLLOWAY, J. P., GROSSKOPF, M. J., KURANZ, C. C., RUTTER, E., TRANTHAM, M. and DRAKE, P. R. (2015). Calibrating a large computer experiment simulating radiative shock hydrodynamics. *Ann. Appl. Stat.* **9** 1141–1168. [MR3418718](#)
- HALL, P. and MAITI, T. (2006). On parametric bootstrap methods for small area prediction. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 221–238. [MR2188983](#)
- HAMMER, A., HEINEMANN, D., LORENZ, E. and LUCKEHE, B. (1999). Short-term forecasting of solar radiation: A statistical approach using satellite data. *Sol. Energy* **67** 139–150.
- HERSBACH, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15** 559–570.
- JIANG, H., SCHÖRGENDORFER, A., HWANG, Y. and AMEMIYA, Y. (2015). A practical approach to spatio-temporal analysis. *Statist. Sinica* **25** 369–384. [MR3328820](#)
- KIM, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika* **98** 119–132. [MR2804214](#)
- KLEIN, L. J., MARIANNO, F. J., ALBRECHT, C. M., FREITAG, M., LU, S., HINDS, N., SHAO, X., RODRIGUEZ, S. B. and HAMANN, H. F. (2015). PAIRS: A scalable geo-spatial data analytics platform. In *2015 IEEE International Conference on Big Data (Big Data)* 1290–1298.
- KRÜGER, F., LERCH, S., THORARINSDOTTIR, T. L. and GNEITING, T. (2016). Probabilistic forecasting and comparative model assessment based on Markov chain Monte Carlo output. Preprint. Available at [arXiv:1608.06802](#).
- LANGE, K. L., LITTLE, R. J. A. and TAYLOR, J. M. G. (1989). Robust statistical modeling using the t distribution. *J. Amer. Statist. Assoc.* **84** 881–896. [MR1134486](#)
- LIU, F., BAYARRI, M. J. and BERGER, J. O. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Anal.* **4** 119–150. [MR2486241](#)
- LIU, X., YEO, K., HWANG, Y., SINGH, J. and KALAGNANAM, J. (2016). A statistical modeling approach for air quality data based on physical dispersion processes and its application to ozone modeling. *Ann. Appl. Stat.* **10** 756–785. [MR3528359](#)
- MARGOLIS, R., COGGESHALL, C. and ZUBOY, J. (2012). Integration of solar into the U.S. electric power system. In *SunShot Vision Study* U.S. Department of Energy, Washington, DC.
- MARQUEZ, R. and COIMBRA, C. F. M. (2013). Intra-hour DNI forecasting based on cloud tracking image analysis. *Sol. Energy* **91** 327–336.
- MATHIESEN, P., COLLIER, C. and KLEISSL, J. (2013). A high-resolution, cloud-assimilating numerical weather prediction model for solar irradiance forecasting. *Sol. Energy* **92** 47–61.
- MATHIESEN, P. and KLEISSL, J. (2011). Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. *Sol. Energy* **85** 967–977.
- ORWIG, K., AHLSTROM, M., BANUNARAYANAN, V., SHARP, J., WILCZAK, J., FREEDMAN, J., HAUPT, S., CLINE, J., BARTHOLOMY, O., HAMANN, H., HODGE, B., FINLEY, C., NAKAFUJI, D., PETERSON, J., MAGGIO, D. and MARQUIS, M. (2015). Recent trends in variable

- generation forecasting and its value to the power system. *IEEE Trans. Sustain. Energy* **6** 924–933.
- PELLAND, S., GALANIS, G. and KALLOS, G. (2013). Solar and photovoltaic forecasting through post-processing of the global environmental multiscale numerical weather prediction model. *Prog. Photovolt.* **21** 284–296.
- PEREZ, R., INEICHEN, P., MOORE, K., KMIECIK, M., CHAIN, C., GEORGE, R. and VIGNOLA, F. (2002). A new operational model for satellite-derived irradiances: Description and validation. *Sol. Energy* **73** 307–317.
- PEREZ, R., LORENZ, E., PELLAND, S., BEAUHARNOIS, M., KNOWE, G. V., HEMKER JR., K., HEINEMANN, D., REMUND, J., MULLER, S. C., TRAUNMULLER, W., STEINMAUER, G., POZO, D., RUIZ-ARIAS, J. A., LARA-FANEGO, V., RAMIREZ-SANTIGOSA, L., GASTON-ROMERO, M. and POMARES, L. M. (2013). Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. *Sol. Energy* **94** 305–326.
- PRATOLA, M. T., CHIPMAN, H. A., GATTIKER, J. R., HIGDON, D. M., MCCULLOCH, R. and RUST, W. N. (2014). Parallel Bayesian additive regression trees. *J. Comput. Graph. Statist.* **23** 830–852. [MR3224658](#)
- QIAN, P. Z. G. and WU, C. F. J. (2008). Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics* **50** 192–204. [MR2439878](#)
- SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer, New York. [MR2160708](#)
- SCOTT, S. L., BLOCKER, A. W., BONASSI, F. V., CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *Int. J. Manag. Sci. Eng. Manag.* **11** 78–88.
- SKAMAROCK, W. C., KLEMP, J. B., DUDHIA, J., GILL, D. O., BARKER, D. M., DUDA, M. G., HUANG, X.-Y., WANG, W. and POWERS, J. G. (2008). A description of the advanced research WRF version 3. NCAR Technical Note: NCAR/TN-475+STR, National Center for Atmospheric Research, Boulder, CO.
- SOTO, W. D., KLEIN, S. A. and BECKMAN, W. A. (2006). Improvement and validation of a model for photovoltaic array performance. *Sol. Energy* **80** 78–88.
- WELCH, W. J., BUCH, R. J., SACKS, J., WYNN, H. P., MITCHELL, T. J. and MORRIS, M. D. (1992). Screening, predicting and computer experiments. *Technometrics* **34** 15–25.
- WONG, G. Y. and MASON, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *J. Amer. Statist. Assoc.* **80** 513–524.
- WONG, R. K. W., STORLIE, C. B. and LEE, T. C. M. (2017). A frequentist approach to computer model calibration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 635–648. [MR3611763](#)
- WU, C. F. J. (2015). Post-Fisherian experimentation: From physical to virtual. *J. Amer. Statist. Assoc.* **110** 612–620. [MR3367251](#)
- ZHANG, J., HODGE, B.-M., LU, S., HAMANN, H. F., LEHMAN, B., SIMMONS, J., CAMPOS, E., BANUNARAYANAN, V., BLACK, J. and TEDESCO, J. (2015a). Baseline and target values for regional and point PV power forecasts: Toward improved solar forecasting. *Sol. Energy* **122** 804–819.
- ZHANG, J., FLORITA, A., HODGE, B.-M., LU, S., HAMANN, H. F., BANUNARAYANAN, V. and BROCKWAY, A. M. (2015b). A suite of metrics for assessing the performance of solar power forecasting. *Sol. Energy* **111** 157–175.

***e*PCA: HIGH DIMENSIONAL EXPONENTIAL FAMILY PCA**

BY LYDIA T. LIU^{*}, EDGAR DOBRIBAN^{†,1,2} AND AMIT SINGER^{‡,3}

University of California at Berkeley^{}, University of Pennsylvania[†]
and Princeton University[‡]*

Many applications involve large datasets with entries from exponential family distributions. Our main motivating application is photon-limited imaging, where we observe images with Poisson distributed pixels. We focus on X-ray Free Electron Lasers (XFEL), a quickly developing technology whose goal is to reconstruct molecular structure. In XFEL, estimating the principal components of the noiseless distribution is needed for denoising and for structure determination. However, the standard method, Principal Component Analysis (PCA), can be inefficient in non-Gaussian noise.

Motivated by this application, we develop *e*PCA (exponential family PCA), a new methodology for PCA on exponential families. *e*PCA is a fast method that can be used very generally for dimension reduction and denoising of large data matrices with exponential family entries.

We conduct a substantive XFEL data analysis using *e*PCA. We show that *e*PCA estimates the PCs of the distribution of images more accurately than PCA and alternatives. Importantly, it also leads to better denoising. We also provide theoretical justification for our estimator, including the convergence rate and the Marchenko–Pastur law in high dimensions. An open-source implementation is [available](#).

REFERENCES

- ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* **11** R106.
- ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. [MR1990662](#)
- ANSCOMBE, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35** 246–254. [MR0028556](#)
- BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. Springer, New York. [MR2567175](#)
- BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. [MR2165575](#)
- BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. [MR2279680](#)
- BARTHOLOMEW, D. J. and KNOTT, M. (1999). *Latent Variable Models and Factor Analysis*, 2nd ed. *Kendall's Library of Statistics* **7**. Edward Arnold, London. [MR1711686](#)
- BASRI, R. and JACOBS, D. W. (2003). Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **25** 218–233.
- BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. [MR2782201](#)

Key words and phrases. Denoising, shrinkage, XFEL imaging, random matrix theory.

- BERGMANN, U., YACHANDRA, V. and YANO, J., eds. (2017). *X-Ray Free Electron Lasers*. The Royal Society of Chemistry, Croydon.
- BHAMRE, T., ZHANG, T. and SINGER, A. (2016). Denoising and covariance estimation of single particle cryo-EM images. *Journal of Structural Biology* **195** 72–81.
- BIGOT, J., DELEDALLE, C. and FÉRAL, D. (2016). Generalized SURE for optimal shrinkage of singular values in low-rank matrix denoising. Preprint. Available at [arXiv:1605.07412](https://arxiv.org/abs/1605.07412).
- BOUCHERON, S., BOUSQUET, O., LUGOSI, G. and MASSART, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.* **33** 514–560. [MR2123200](#)
- CAO, Y. and XIE, Y. (2014). Low-rank matrix recovery in Poisson noise. In *Signal and Information Processing (GlobalSIP)*, 2014 *IEEE Global Conference on* 384–388. IEEE, New York.
- CHEN, X. and STOREY, J. D. (2015). Consistent estimation of low-dimensional latent structure in high-dimensional data. Preprint. Available at [arXiv:1510.03497](https://arxiv.org/abs/1510.03497).
- COLLINS, M., DASGUPTA, S. and SCHAPIRE, R. (2001). A generalization of principal component analysis to the exponential family. *Advances in Neural Information Processing Systems (NIPS)*.
- DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. and HARSHMAN, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science* **41** 391–407.
- DOBRIAN, E. (2015). Efficient computation of limit spectra of sample covariance matrices. *Random Matrices Theory Appl.* **4** 1550019, 36. [MR3418848](#)
- DOBRIAN, E. (2017). Sharp detection in PCA under correlations: All eigenvalues matter. *Ann. Statist.* **45** 1810–1833. [MR3670197](#)
- DONOHO, D., GAVISH, M. and JOHNSTONE, I. (2013). Optimal shrinkage of eigenvalues in the spiked covariance model. Preprint. Available at [arXiv:1311.0851](https://arxiv.org/abs/1311.0851).
- FAVRE-NICOLIN, V., BARUCHEL, J., RENEVIER, H., EYMERY, J. and BORBÉLY, A. (2015). XTOP: High-resolution X-ray diffraction and imaging. *Journal of Applied Crystallography* **48** 620–620.
- FREEMAN, M. F. and TUKEY, J. W. (1950). Transformations related to the angular and the square root. *Ann. Math. Stat.* **21** 607–611. [MR0038028](#)
- FURNIVAL, T., LEARY, R. K. and MIDGLEY, P. A. (2017). Denoising time-resolved microscopy image sequences with singular value thresholding. *Ultramicroscopy* **178** 112–124.
- HANTKE, M. F., EKEBERG, T. and MAIA, F. R. N. C. (2016). Condor: A simulation tool for Flash X-Ray imaging. *Journal of Applied Crystallography* **49** 1356–1362.
- HUBER, P., RONCHETTI, E. and VICTORIA-FESER, M.-P. (2004). Estimation of generalized linear latent variable models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 893–908. [MR2102471](#)
- JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. [MR1863961](#)
- JOLLIFFE, I. T. (2002). *Principal Component Analysis*, 2nd ed. Springer, New York. [MR2036084](#)
- JOSSE, J. and WAGER, S. (2016). Bootstrap-based regularization for low-rank matrix estimation. *J. Mach. Learn. Res.* **17** 1–29. [MR3555015](#)
- KAM, Z. (1977). Determination of macromolecular structure in solution by spatial correlation of scattering fluctuations. *Macromolecules* **10** 927–934.
- KAM, Z. (1980). The reconstruction of structure from electron micrographs of randomly oriented particles. *J. Theoret. Biol.* **82** 15–39.
- KURTA, R. P., DONATELLI, J. J., YOON, C. H. et al. (2017). Correlations in scattered X-Ray laser pulses reveal nanoscale structural features of viruses. *Phys. Rev. Lett.* **119** 158102.
- LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. [MR2026339](#)
- LEE, S., ZOU, F. and WRIGHT, F. A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *Ann. Statist.* **38** 3605–3629. [MR2766862](#)
- LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. [MR2135927](#)

- LI, J. and TAO, D. (2010). Simple exponential family PCA. In *AISTATS* 453–460.
- LIU, L. T., DOBRIBAN, E. and SINGER, A. (2018). Supplement to “ePCA: High dimensional exponential family PCA.” DOI:[10.1214/18-AOAS1146SUPP](https://doi.org/10.1214/18-AOAS1146SUPP).
- MAIA, F. R. N. C. and HAJDU, J. (2016). The trickle before the torrent-diffraction data from X-ray lasers. *Sci. Data* **3** 160059.
- MÄKITALO, M. and FOI, A. (2011). Optimal inversion of the Anscombe transformation in low-count Poisson image denoising. *IEEE Trans. Image Process.* **20** 99–109. [MR2767607](#)
- MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb.* **72** 507–536. [MR0208649](#)
- MARTIN, A. V., WANG, F., LOH, N. D., EKEBERG, T. et al. (2012). Noise-robust coherent diffractive imaging with a single diffraction pattern. *Opt. Express* **20** 16650–16661.
- NADAKUDITI, R. R. (2014). OptShrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Trans. Inform. Theory* **60** 3002–3018. [MR3200641](#)
- NOWAK, R. D. and BARANIUK, R. G. (1999). Wavelet-domain filtering for photon imaging systems. *IEEE Trans. Image Process.* **8** 666–678.
- PANDE, K., SCHWANDER, P., SCHMIDT, M. and SALDIN, D. (2014). Deducing fast electron density changes in randomly orientated uncrystallized biomolecules in a pump–probe experiment. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **369** 20130332.
- PANDE, K., SCHMIDT, M., SCHWANDER, P. and SALDIN, D. K. (2015). Simulations on time-resolved structure determination of uncrystallized biomolecules in the presence of shot noise. *Struct. Dyn.* **2** 024103.
- PATTERSON, N., PRICE, A. L. and REICH, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* **2** e190.
- PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. [MR2399865](#)
- SALDIN, D. K., SHNEERSON, V. L., FUNG, R. and OURMAZD, A. (2009). Structure of isolated biomolecules obtained from ultrashort x-ray pulses: Exploiting the symmetry of random orientations. *J. Phys., Condens. Matter* **21** 134014.
- SCHWANDER, P., GIANNAKIS, D., YOON, C. H. and OURMAZD, A. (2012). The symmetries of image formation by scattering. II. Applications. *Opt. Express* **20** 12827–12849.
- SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. Wiley, New York. [MR1190470](#)
- SHABALIN, A. A. and NOBEL, A. B. (2013). Reconstruction of a low-rank matrix in the presence of Gaussian noise. *J. Multivariate Anal.* **118** 67–76. [MR3054091](#)
- STARCK, J.-L., MURTAGH, F. and FADILI, J. M. (2010). *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*. Cambridge Univ. Press, Cambridge. [MR2643260](#)
- STARODUB, D., AQUILA, A., BAJT, S. et al. (2012). Single-particle structure determination by correlations of snapshot X-ray diffraction patterns. *Nat. Commun.* **3**.
- STEGLE, O., TEICHMANN, S. A. and MARIONI, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16** 133–145.
- STEIN, C. (1956). Some problems in multivariate analysis. Technical report, Dept. Statistics, Stanford Univ., Stanford, CA.
- TROPP, J. A. (2016). The expected norm of a sum of independent random matrices: An elementary approach. In *High Dimensional Probability VII. Progress in Probability* **71** 173–202. Springer, Cham. [MR3565264](#)
- UDELL, M., HORN, C., ZADEH, R. and BOYD, S. (2014). Generalized low rank models. In *NIPS Workshop on Distributed Machine Learning and Matrix Computations*.
- UDELL, M., HORN, C., ZADEH, R. and BOYD, S. (2016). Generalized low rank models. *Found. Trends Mach. Learn.* **9** 1–118.

- VISSCHER, P. M., BROWN, M. A., MCCARTHY, M. I. and YANG, J. (2012). Five years of GWAS discovery. *Am. J. Hum. Genet.* **90** 7–24.
- YAO, J., ZHENG, S. and BAI, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis. Cambridge Series in Statistical and Probabilistic Mathematics* **39**. Cambridge Univ. Press, New York. [MR3468554](#)
- ZHAO, Z., SHKOLNISKY, Y. and SINGER, A. (2016). Fast steerable principal component analysis. *IEEE Trans. Comput. Imaging* **2** 1–12. [MR3472531](#)

RANK TESTS IN UNMATCHED CLUSTERED RANDOMIZED TRIALS APPLIED TO A STUDY OF TEACHER TRAINING

BY PENG DING¹ AND LUKE KEELE

University of California, Berkeley and Georgetown University

In the Teacher and Leader Performance Evaluation Systems study, schools were randomly assigned to receive new measures of teacher and principal performance. One outcome in the study, measured at the teacher level, was truncated at zero, and displayed a long tail. Rank-based statistics are one natural method to apply to such outcomes, since inferences will be robust and exact, and we can avoid assumptions about the model that generated the data. We investigate four different possible rank statistics that vary in the form of weighting applied to clusters. Each test statistic has the correct level but may vary in terms of the power to detect departures from the null. We conduct simulations for power comparing to linear mixed models with Normal, t , and Cauchy errors. We obtain a point estimate and construct confidence intervals by applying the Tobit model of effects, which assumes that treatment increases the outcome by a constant amount but only if the response under control would be positive. We also develop a formal randomization-based method for testing the appropriateness of the Tobit model of effects. In the data from the study, we find no evidence against the Tobit model of effects.

REFERENCES

- ARONOW, P. M., MIDDLETON, J. A. et al. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference* **1** 135–154.
- BERGER, R. L. and BOOS, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* **89** 1012–1016. [MR1294746](#)
- BORMAN, G. D., SLAVIN, R. E., CHEUNG, A., CHAMBERLAIN, A. M., MADDEN, N. A. and CHAMBERS, B. (2005). Success for all: First-year results from the national randomized field trial. *Educ. Eval. Policy Anal.* **27** 1–22.
- BRAUN, T. M. and FENG, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. *J. Amer. Statist. Assoc.* **96** 1424–1432. [MR1946587](#)
- CHETTY, R., FRIEDMAN, J. N. and ROCKOFF, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *Am. Econ. Rev.* **104** 2593–2632.
- COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York. [MR0474575](#)
- CORNFIELD, J. (1978). Randomization by group. *Am. J. Epidemiol.* **108** 100–102.
- DATTA, S. and SATTEN, G. A. (2005). Rank-sum tests for clustered data. *J. Amer. Statist. Assoc.* **100** 908–915. [MR2201018](#)
- DING, P. (2017). A paradox from randomization-based causal inference. *Statist. Sci.* **32** 331–345. [MR3695995](#)
- DING, P., FELLER, A. and MIRATRIX, L. (2016). Randomization inference for treatment effect variation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 655–671. [MR3506797](#)

Key words and phrases. Fisher randomization test, clustered randomization, model checking, rank statistic, Tobit model.

- DONNER, A. and KLAR, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Wiley, New York.
- DUTTA, S. and DATTA, S. (2016). A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative. *Biometrics* **72** 432–440. [MR3515770](#)
- FEDER, G., GRIFFITHS, C., ELDRIDGE, S. and SPENCE, M. (1999). Effect of postal prompts to patients and general practitioners on the quality of primary care after a coronary event (POST): Randomised controlled trial. *BMJ* **318** 1522–1526.
- FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, London.
- GAIL, M. H., BYAR, D. P., PECHACEK, T. F., CORLE, D. K., GROUP, C. S. et al. (1992). Aspects of statistical design for the community intervention trial for smoking cessation (COMMIT). *Control. Clin. Trials* **13** 6–21.
- GAIL, M. H., MARK, S. D., CARROLL, R. J. and GREEN, S. B. (1996). On design considerations and randomization-based inference for community intervention trials. *Stat. Med.* **15** 1069–1092.
- HANSEN, B. B. and BOWERS, J. (2009). Attributing effects to a cluster-randomized get-out-the-vote campaign. *J. Amer. Statist. Assoc.* **104** 873–885. [MR2562000](#)
- HAYES, R. and MOULTON, L. (2009). *Cluster Randomised Trials*. Chapman& Hall/CRC, London.
- HEDGES, L. V. and HEDBERG, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educ. Eval. Policy Anal.* **29** 60–87.
- HODGES, J. L. JR. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Stat.* **34** 598–611. [MR0152070](#)
- IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—For Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. [MR3309951](#)
- IMBENS, G. M. and WOOLDRIDGE, J. M. (2008). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* **47** 5–86.
- LI, X. and DING, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.* **112** 1759–1769. [MR3750897](#)
- MIDDLETON, J. A. (2008). Bias of the regression estimator for experiments using clustered random assignment. *Statist. Probab. Lett.* **78** 2654–2659. [MR2542462](#)
- MIDDLETON, J. A. and ARONOW, P. M. (2015). Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy* **6** 39–75.
- MURNANE, R. J. and WILLETT, J. B. (2010). *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. Oxford University Press, Oxford.
- NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. Dąbrowska and T. P. Speed. [MR1092986](#)
- NEYMAN, J. (1935). Statistical problems in agricultural experimentation. *Suppl. J. R. Stat. Soc.* **2** 107–180.
- NOLEN, T. L. and HUDGENS, M. G. (2011). Randomization-based inference within principal strata. *J. Amer. Statist. Assoc.* **106** 581–593. [MR2847972](#)
- ROSENBAUM, P. R. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika* **88** 219–231. [MR1841270](#)
- ROSENBAUM, P. R. (2002a). *Observational Studies*, 2nd ed. Springer, New York. [MR1899138](#)
- ROSENBAUM, P. R. (2002b). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. [MR1962487](#)
- ROSENBAUM, P. R. (2007). Confidence intervals for uncommon but dramatic responses to treatment. *Biometrics* **63** 1164–1171, 1313. [MR2414594](#)
- ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York. [MR2561612](#)
- ROSNER, B., GLYNN, R. J. and LEE, M.-L. T. (2003). Incorporation of clustering effects for the Wilcoxon rank sum test: A large-sample approach. *Biometrics* **59** 1089–1098. [MR2025134](#)
- ROSNER, B., GLYNN, R. J. and LEE, M.-L. T. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics* **62** 185–192, 318. [MR2226572](#)

- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **6** 688–701.
- RUBIN, D. B. (1986). Which ifs have causal answers. *J. Amer. Statist. Assoc.* **81** 961–962.
- SCHOCHET, P. Z. (2013). Estimators for clustered education RCTs using the Neyman model for causal inference. *J. Educ. Behav. Stat.* **38** 219–238.
- SMALL, D. S., HAVE, T. R. T. and ROSENBAUM, P. R. (2008). Randomization inference in a group-randomized trial of treatments for depression: Covariate adjustment, noncompliance, and quantile effects. *J. Amer. Statist. Assoc.* **103** 271–279. [MR2420232](#)
- TOBIN, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26** 24–36. [MR0090462](#)
- WAYNE, J. A., GARET, M. S., BROWN, S., RICKLES, J., SONG, M. and MANZESKE, D. (2016). Early Implementation Findings From a Study of Teacher and Principal Performance Measurement and Feedback Year 1 Report. Technical report, American Institutes of Research, Washington, DC.
- WILLIAMSON, J. M., DATTA, S. and SATTEN, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59** 36–42. [MR1978471](#)
- ZHANG, K., TRASKIN, M. and SMALL, D. S. (2012). A powerful and robust test statistic for randomization inference in group-randomized trials with matched pairs of groups. *Biometrics* **68** 75–84. [MR2909855](#)

MARKED SELF-EXCITING POINT PROCESS MODELLING OF INFORMATION DIFFUSION ON TWITTER

BY FENG CHEN AND WAI HONG TAN

UNSW Sydney

Information diffusion occurs on microblogging platforms like Twitter as retweet cascades. When a tweet is posted, it may be retweeted and henceforth further retweeted, and the retweeting process continues iteratively and indefinitely. A natural measure of the popularity of a tweet is the number of retweets it generates. Accurate predictions of tweet popularity can assist Twitter to rank contents more effectively and facilitate the assessment of potential for marketing and campaigning strategies. In this paper, we propose a model called the Marked Self-Exciting Process with Time-Dependent Excitation Function, or MaSEPTiDE for short, to model the retweeting dynamics and to predict the tweet popularity. Our model does not require expensive feature engineering but is capable of leveraging the observed dynamics to accurately predict the future evolution of retweet cascades. We apply our proposed methodology on a large amount of Twitter data and report substantial improvement in prediction performance over existing approaches in the literature.

REFERENCES

- AGARWAL, D., CHEN, B.-C. and ELANGO, P. (2009). Spatio-temporal models for estimating click-through rate. In *Proceedings of the 18th International Conference on World Wide Web* 21–30. ACM, New York.
- AHMED, M., SPAGNA, S., HUICI, F. and NICCOLINI, S. (2013). A peek into the future: Predicting the evolution of popularity in user generated content. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* 607–616. ACM, New York.
- ALVES, R. A., ASSUNÇÃO, R. and DE MELO, P. O. (2016). Burstiness scale: A highly parsimonious model for characterizing random series of events. Preprint. Available at [arXiv:1602.06431](https://arxiv.org/abs/1602.06431).
- BAKSHY, E., HOFMAN, J. M., MASON, W. A. and WATTS, D. J. (2011). Everyone’s an influencer: Quantifying influence on Twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* 65–74. ACM, New York.
- BARABASI, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature* **435** 207–211.
- CHA, M., HADDADI, H., BENEVENUTO, F. and GUMMADI, P. K. (2010). Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM-2010)* 10–17. AAAI Press, Palo Alto, CA.
- CHEN, F. and HALL, P. (2013). Inference for a nonstationary self-exciting point process with an application in ultra-high frequency financial data modeling. *J. Appl. Probab.* **50** 1006–1024. [MR3161370](https://doi.org/10.1214/12-AJP1170)

Key words and phrases. B-spline, forecast, Hawkes process, integral equation, nonstationary self-exciting point process, popularity prediction, simulation.

- CHEN, F. and HALL, P. (2016). Nonparametric estimation for self-exciting point processes—A parsimonious approach. *J. Comput. Graph. Statist.* **25** 209–224. [MR3474044](#)
- CHEN, F. and STINDL, T. (2018). Direct likelihood evaluation for the renewal Hawkes process. *J. Comput. Graph. Statist.* **27** 119–131.
- CRANE, R. and SORNETTE, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci. USA* **105** 15649–15653.
- DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes, Vol. I: Elementary Theory and Methods*, 2nd ed. Springer, New York. [MR1950431](#)
- FOX, E. W., SHORT, M. B., SCHOENBERG, F. P., CORONGES, K. D. and BERTOZZI, A. L. (2016). Modeling e-mail networks and inferring leadership using self-exciting point processes. *J. Amer. Statist. Assoc.* **111** 564–584. [MR3538687](#)
- GAO, S., MA, J. and CHEN, Z. (2015). Modeling and predicting retweeting dynamics on microblogging platforms. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* 107–116. ACM, New York.
- GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. [MR2847988](#)
- HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. [MR0278410](#)
- KOBAYASHI, R. and LAMBIOTTE, R. (2016). TiDeH: Time-dependent Hawkes process for predicting retweet dynamics. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM-2016)* 191–200. The AAAI Press, Palo Alto, CA.
- 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.
- LEWIS, P. A. W. and SHEDLER, G. S. (1979). Simulation of nonhomogeneous Poisson processes by thinning. *Nav. Res. Logist. Q.* **26** 403–413. [MR0546120](#)
- LI, C.-T., SHAN, M.-K., JHENG, S.-H. and CHOU, K.-C. (2016). Exploiting concept drift to predict popularity of social multimedia in microblogs. *Inform. Sci.* **339** 310–331.
- LYMPEROPOULOS, I. N. (2016). Predicting the popularity growth of online content: Model and algorithm. *Inform. Sci.* **369** 585–613.
- MATSUBARA, Y., SAKURAI, Y., PRAKASH, B. A., LI, L. and FALOUTSOS, C. (2012). Rise and fall patterns of information diffusion: Model and implications. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 6–14. ACM, New York.
- MISHRA, S., RIZOIU, M.-A. and XIE, L. (2016). Feature driven and point process approaches for popularity prediction. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* 1069–1078. ACM, New York.
- NAVEED, N., GOTTRON, T., KUNEGIS, J. and ALHADI, A. C. (2011). Bad news travel fast: A content-based analysis of interestingness on Twitter. In *Proceedings of the 3rd International Web Science Conference* Art. ID 8. ACM, New York.
- NELDER, J. A. and MEAD, R. (1965). A simplex method for function minimization. *Comput. J.* **7** 308–313. [MR3363409](#)
- OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.
- R CORE TEAM (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.
- SZABO, G. and HUBERMAN, B. A. (2010). Predicting the popularity of online content. *Commun. ACM* **53** 80–88.
- TUMASIAN, A., SPRENGER, T. O., SANDNER, P. G. and WELPE, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth*

- International Conference on Weblogs and Social Media (ICWSM-2010)* 178–185. AAAI Press, Palo Alto, CA.
- WU, B., CHENG, W.-H., ZHANG, Y. and MEI, T. (2016). Time matters: Multi-scale temporalization of social media popularity. In *Proceedings of the 2016 ACM on Multimedia Conference* 1336–1344. ACM, New York.
- YAN, Y., TAN, Z., GAO, X., TANG, S. and CHEN, G. (2016). STH-Bass: A spatial-temporal heterogeneous bass model to predict single-tweet popularity. In *International Conference on Database Systems for Advanced Applications* 18–32. Springer, Cham.
- ZAMAN, T., FOX, E. B. and BRADLOW, E. T. (2014). A Bayesian approach for predicting the popularity of tweets. *Ann. Appl. Stat.* **8** 1583–1611. [MR3271345](#)
- ZHAO, Q., ERDOGDU, M. A., HE, H. Y., RAJARAMAN, A. and LESKOVEC, J. (2015). SEISMIC: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1513–1522. ACM, New York.

STANDARDIZATION OF MULTIVARIATE GAUSSIAN MIXTURE MODELS AND BACKGROUND ADJUSTMENT OF PET IMAGES IN BRAIN ONCOLOGY¹

BY MENG LI AND ARMIN SCHWARTZMAN

Rice University and University of California, San Diego

In brain oncology, it is routine to evaluate the progress or remission of the disease based on the differences between a pre-treatment and a post-treatment Positron Emission Tomography (PET) scan. Background adjustment is necessary to reduce confounding by tissue-dependent changes not related to the disease. When modeling the voxel intensities for the two scans as a bivariate Gaussian mixture, background adjustment translates into standardizing the mixture at each voxel, while tumor lesions present themselves as outliers to be detected. In this paper, we address the question of how to standardize the mixture to a standard multivariate normal distribution, so that the outliers (i.e., tumor lesions) can be detected using a statistical test. We show theoretically and numerically that the tail distribution of the standardized scores is favorably close to standard normal in a wide range of scenarios while being conservative at the tails, validating voxelwise hypothesis testing based on standardized scores. To address standardization in spatially heterogeneous image data, we propose a spatial and robust multivariate expectation-maximization (EM) algorithm, where prior class membership probabilities are provided by transformation of spatial probability template maps and the estimation of the class mean and covariances are robust to outliers. Simulations in both univariate and bivariate cases suggest that standardized scores with soft assignment have tail probabilities that are either very close to or more conservative than standard normal. The proposed methods are applied to a real data set from a PET phantom experiment, yet they are generic and can be used in other contexts.

REFERENCES

- ASHBURNER, J. (2012). SPM: A history. *NeuroImage* **62** 791–800.
 ASHBURNER, J. and FRISTON, K. J. (2005). Unified segmentation. *NeuroImage* **26** 839–851.
 BAI, B., BADING, J. and CONTI, P. S. (2013). Tumor quantification in clinical positron emission tomography. *Theranostics* **3** 787–801.
 BESAG, J. (1986). On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B* **48** 259–302. [MR0876840](#)
 BORGHAMMER, P., AANERUD, J. and GJEDDE, A. (2009). Data-driven intensity normalization of PET group comparison studies is superior to global mean normalization. *NeuroImage* **46** 981–988.
 CAMPBELL, N. A. (1984). Mixture models and atypical values. *Math. Geol.* **16** 465–477.

Key words and phrases. Background adjustment, PET images, tumor detection, multivariate Gaussian mixture model, outlier detection, robust EM algorithm, spatial modeling, standardized scores, tail distributions, voxelwise inference.

- CHEN, J. L., GUNN, S. R., NIXON, M. S. and GUNN, R. N. (2001). Markov random field models for segmentation of PET images. In *Biennial International Conference on Information Processing in Medical Imaging* 468–474. Springer, Berlin.
- DASGUPTA, A., HOPCROFT, J., KLEINBERG, J. and SANDLER, M. (2005). On learning mixtures of heavy-tailed distributions. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)* 491–500. IEEE, New York.
- 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](#)
- DEVLIN, S. J., GNANADESIKAN, R. and KETTENRING, J. R. (1981). Robust estimation of dispersion matrices and principal components. *J. Amer. Statist. Assoc.* **76** 354–362.
- FIGUEIREDO, M. A. T. and JAIN, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 381–396.
- GUO, M., YAP, J. T., DEN ABEELE, A. D., LIN, N. U. and SCHWARTZMAN, A. (2014). Voxelwise single-subject analysis of imaging metabolic response to therapy in neuro-oncology. *Stat.* **3** 172–186.
- GUPTA, M. R. and CHEN, Y. (2011). *Theory and Use of the EM Algorithm*. Now Publishers Inc.
- HANSON, T. E. (2006). Inference for mixtures of finite Polya tree models. *J. Amer. Statist. Assoc.* **101** 1548–1565. [MR2279479](#)
- HOFFMAN, E. J., CUTLER, P. D., GUERRERO, T. M., DIGBY, W. M. and MAZZIOTTA, J. C. (1991). Assessment of accuracy of PET utilizing a 3-D phantom to simulate the activity distribution of [18F] fluorodeoxyglucose uptake in the human brain. *J. Cereb. Blood Flow Metab.* **11** A17–A25.
- HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. [MR0161415](#)
- LEAHY, R. M. and QI, J. (2000). Statistical approaches in quantitative positron emission tomography. *Stat. Comput.* **10** 147–165.
- LEE, Y.-Y., CHOI, C. H., KIM, C. J., KANG, H., KIM, T.-J., LEE, J.-W., LEE, J.-H., BAE, D.-S. and KIM, B.-G. (2009). The prognostic significance of the SUVmax (maximum standardized uptake value for F-18 fluorodeoxyglucose) of the cervical tumor in PET imaging for early cervical cancer: Preliminary results. *Gynecol. Oncol.* **115** 65–68.
- LI, M. and SCHWARTZMAN, A. (2018). Supplement to “Standardization of multivariate Gaussian mixture models and background adjustment of PET images in brain oncology.” DOI:[10.1214/18-AOAS1149SUPP](#).
- LIN, T. I., LEE, J. C. and YEN, S. Y. (2007). Finite mixture modelling using the skew normal distribution. *Statist. Sinica* **17** 909–927. [MR2408641](#)
- LO, Y., MENDELL, N. R. and RUBIN, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika* **88** 767–778. [MR1859408](#)
- MARONNA, R. A. (1976). Robust M -estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67. [MR0388656](#)
- MARONNA, R. A., MARTIN, R. D. and YOHAI, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, Chichester. [MR2238141](#)
- MCLACHLAN, G. J. and BASFORD, K. E. (1988). *Mixture Models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs* **84**. Dekker, New York. [MR0926484](#)
- MCLACHLAN, G. J. and KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2nd ed. Wiley, Hoboken, NJ. [MR2392878](#)
- MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. Wiley-Interscience, New York. [MR1789474](#)
- NGUYEN, T. M. and WU, Q. M. J. (2012). Gaussian-mixture-model-based spatial neighborhood relationships for pixel labeling problem. *IEEE Trans. Syst. Man Cybern., Part B, Cybern.* **42** 193–202.

- O'SULLIVAN, F., MUZI, M., MANKOFF, D. A., EARY, J. F., SPENCE, A. M. and KROHN, K. A. (2014). Voxel-level mapping of tracer kinetics in PET studies: A statistical approach emphasizing tissue life tables. *Ann. Appl. Stat.* **8** 1065–1094. [MR3262546](#)
- PEEL, D. and MCLACHLAN, G. J. (2000). Robust mixture modelling using the t distribution. *Stat. Comput.* **10** 339–348.
- QIN, Y. and PRIEBE, C. E. (2013). Maximum L_q -likelihood estimation via the expectation-maximization algorithm: A robust estimation of mixture models. *J. Amer. Statist. Assoc.* **108** 914–928. [MR3174673](#)
- QIN, L., SCHWARTZMAN, A., MCCALL, K., KACHOUIE, N. N. and YAP, J. T. (2017). Method for detecting voxelwise changes in fluorodeoxyglucose-positron emission tomography brain images via background adjustment in cancer clinical trials. *J. Med. Imag.* **4** 024006.
- REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. [MR0738930](#)
- 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](#)
- SANJAY-GOPAL, S. and HEBERT, T. (1998). Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm. *IEEE Trans. Image Process.* **7** 1014–1028.
- SOFFIENTINI, C. D., DE BERNARDI, E., ZITO, F., CASTELLANI, M. and BASELLI, G. (2016). Background based Gaussian mixture model lesion segmentation in PET. *Med. Phys.* **43** 2662–2675.
- SORET, M., BACHARACH, S. L. and BUVAT, I. (2007). Partial-volume effect in PET tumor imaging. *J. Nucl. Med.* **48** 932–945.
- STEPHENS, M. (2000). Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods. *Ann. Statist.* **28** 40–74. [MR1762903](#)
- TAKEDA, A., YOKOSUKA, N., OHASHI, T., KUNIEDA, E., FUJII, H., AOKI, Y., SANUKI, N., KOIKE, N. and OZAWA, Y. (2011). The maximum standardized uptake value (SUVmax) on FDG-PET is a strong predictor of local recurrence for localized non-small-cell lung cancer after stereotactic body radiotherapy (SBRT). *Radiother. Oncol.* **101** 291–297.
- VALK, P. E., BAILEY, D. L., TOWNSEND, D. W. and MAISEY, M. N. (2003). *Positron Emission Tomography: Basic Science and Clinical Practice*. Springer, London.
- VEHTARI, A. and OJANEN, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Stat. Surv.* **6** 142–228. [MR3011074](#)
- VENTURINI, S., DOMINICI, F. and PARMIGIANI, G. (2008). Gamma shape mixtures for heavy-tailed distributions. *Ann. Appl. Stat.* **2** 756–776. [MR2524355](#)
- VLASSIS, N. and LIKAS, A. (1999). A kurtosis-based dynamic approach to Gaussian mixture modeling. *IEEE Trans. Syst. Man Cybern., Part A, Syst. Humans* **29** 393–399.
- WAHL, R. L., JACENE, H., KASAMON, Y. and LODGE, M. A. (2009). From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors. *J. Nucl. Med.* **50** 122S–150S.
- YOUNG, H., BAUM, R., CREMERIUS, U., HERHOLZ, K., HOEKSTRA, O., LAMMERTSMA, A. A., PRUIM, J., PRICE, P. and OTHERS (1999). Measurement of clinical and subclinical tumour response using [18 F]-fluorodeoxyglucose and positron emission tomography: Review and 1999 EORTC recommendations. *Eur. J. Cancer* **35** 1773–1782.
- ZASADNY, K. R. and WAHL, R. L. (1993). Standardized uptake values of normal tissues at PET with 2-[fluorine-18]-fluoro-2-deoxy-D-glucose: Variations with body weight and a method for correction. *Radiology* **189** 847–850.
- ZHANG, J., MODESTINO, J. W. and LANGAN, D. A. (1994). Maximum-likelihood parameter estimation for unsupervised stochastic model-based image segmentation. *IEEE Trans. Image Process.* **3** 404–420.

GAUSSIAN PROCESS MODELLING IN APPROXIMATE BAYESIAN COMPUTATION TO ESTIMATE HORIZONTAL GENE TRANSFER IN BACTERIA

BY MARKO JÄRVENPÄÄ*, MICHAEL U. GUTMANN†, AKI VEHTARI* AND
PEKKA MARTTINEN*,¹

Aalto University and University of Edinburgh†*

Approximate Bayesian computation (ABC) can be used for model fitting when the likelihood function is intractable but simulating from the model is feasible. However, even a single evaluation of a complex model may take several hours, limiting the number of model evaluations available. Modelling the discrepancy between the simulated and observed data using a Gaussian process (GP) can be used to reduce the number of model evaluations required by ABC, but the sensitivity of this approach to a specific GP formulation has not yet been thoroughly investigated. We begin with a comprehensive empirical evaluation of using GPs in ABC, including various transformations of the discrepancies and two novel GP formulations. Our results indicate the choice of GP may significantly affect the accuracy of the estimated posterior distribution. Selection of an appropriate GP model is thus important. We formulate expected utility to measure the accuracy of classifying discrepancies below or above the ABC threshold, and show that it can be used to automate the GP model selection step. Finally, based on the understanding gained with toy examples, we fit a population genetic model for bacteria, providing insight into horizontal gene transfer events within the population and from external origins.

REFERENCES

- ANSARI, M. A. and DIDELOT, X. (2014). Inference of the properties of the recombination process from whole bacterial genomes. *Genetics* **196** 253–265.
- BEAUMONT, M. A., ZHANG, W. and BALDING, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162** 2025–2035.
- BEAUMONT, M. A., CORNUET, J.-M., MARIN, J.-M. and ROBERT, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika* **96** 983–990. [MR2767283](#)
- BERNARDO, J.-M. and SMITH, A. F. M. (2001). *Bayesian Theory*. Wiley, Chichester. [MR1274699](#)
- BLUM, M. G. B. (2010). Approximate Bayesian computation: A nonparametric perspective. *J. Amer. Statist. Assoc.* **105** 1178–1187. [MR2752613](#)
- BLUM, M. G. B. and FRANÇOIS, O. (2010). Non-linear regression models for approximate Bayesian computation. *Stat. Comput.* **20** 63–73. [MR2578077](#)
- BROCHU, E., CORA, V. M. and DE FREITAS, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Preprint. Available at [arXiv:1012.2599](#).

Key words and phrases. Approximate Bayesian computation, intractable likelihood, Gaussian process, input-dependent noise, model selection.

- CHEWAPREECHA, C., HARRIS, S. R., CROUCHER, N. J., TURNER, C., MARTTINEN, P., CHENG, L., PESSIA, A., AANENSEN, D. M., MATHER, A. E., PAGE, A. J. et al. (2014). Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* **46** 305–309.
- COHAN, F. M. and PERRY, E. B. (2007). A systematics for discovering the fundamental units of bacterial diversity. *Curr. Biol.* **17** R373–R386.
- CROUCHER, N. J., HARRIS, S. R., FRASER, C., QUAIL, M. A., BURTON, J., VAN DER LINDEN, M., MCGEE, L., VON GOTTEBERG, A., SONG, J. H., KO, K. S. et al. (2011). Rapid pneumococcal evolution in response to clinical interventions. *Science* **331** 430–434.
- CROUCHER, N. J., FINKELSTEIN, J. A., PELTON, S. I., MITCHELL, P. K., LEE, G. M., PARKHILL, J., BENTLEY, S. D., HANAGE, W. P. and LIPSITCH, M. (2013). Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.* **45** 656–663.
- DEL MORAL, P., DOUCET, A. and JASRA, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat. Comput.* **22** 1009–1020. [MR2950081](#)
- DOROGHAZI, J. R. and BUCKLEY, D. H. (2011). A model for the effect of homologous recombination on microbial diversification. *Genome Biol. Evol.* **3** 1349–1356.
- DROVANDI, C. C., MOORES, M. T. and BOYS, R. J. (2018). Accelerating pseudo-marginal MCMC using Gaussian processes. *Comput. Statist. Data Anal.* **118** 1–17. [MR3715260](#)
- DROVANDI, C. C. and PETTITT, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics* **67** 225–233. [MR2898834](#)
- DROVANDI, C. C., PETTITT, A. N. and LEE, A. (2015). Bayesian indirect inference using a parametric auxiliary model. *Statist. Sci.* **30** 72–95. [MR3317755](#)
- FAN, Y., NOTT, D. J. and Sisson, S. A. (2013). Approximate Bayesian computation via regression density estimation. *Stat* **2** 34–48.
- FEARNHEAD, P. and PRANGLE, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 419–474. [MR2925370](#)
- FRASER, C., HANAGE, W. P. and SPRATT, B. G. (2007). Recombination and the nature of bacterial speciation. *Science* **315** 476–480.
- GOLDBERG, P. W., WILLIAMS, C. K. I. and BISHOP, C. M. (1997). Regression with input-dependent noise: A Gaussian process treatment. *Adv. Neural Inf. Process. Syst.* **10** 493–499.
- GUTMANN, M. U. and CORANDER, J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *J. Mach. Learn. Res.* **17** Paper No. 125, 47. [MR3555016](#)
- HARTIG, F., CALABRESE, J. M., REINEKING, B., WIEGAND, T. and HUTH, A. (2011). Statistical inference for stochastic simulation models – Theory and application. *Ecol. Lett.* **14** 816–827.
- JABOT, F., LAGARRIGUES, G., COURBAUD, B. and DUMOULIN, N. (2014). A comparison of emulation methods for approximate Bayesian computation. Preprint. Available at [arXiv:1412.7560](#).
- JÄRVENPÄÄ, M., GUTMANN, M., PLESKA, A., VEHTARI, A. and MARTTINEN, P. (2017). Efficient acquisition rules for model-based approximate Bayesian computation. Preprint. Available at [arXiv:1704.00520](#).
- JÄRVENPÄÄ, M., GUTMANN, M., VEHTARI, A. and MARTTINEN, P. (2018). Supplement to “Gaussian process modeling in approximate Bayesian computation to estimate horizontal gene transfer in bacteria.” DOI:[10.1214/18-AOAS1150SUPP](#).
- KANDASAMY, K., SCHNEIDER, J. and PÓCZOS, B. (2015). Bayesian active learning for posterior estimation. In *International Joint Conference on Artificial Intelligence* 3605–3611.
- LENORMAND, M., JABOT, F. and DEFFUANT, G. (2013). Adaptive approximate Bayesian computation for complex models. *Comput. Statist.* **28** 2777–2796. [MR3141363](#)
- LINTUSAARI, J., GUTMANN, M. U., DUTTA, R., KASKI, S. and CORANDER, J. (2016). Fundamentals and recent developments in approximate Bayesian computation. *Syst. Biol.* **66** e66–e82.
- MAJEWSKI, J. (2001). Sexual isolation in bacteria. *FEMS Microbiol. Lett.* **199** 161–169.
- MARIN, J.-M., PUDLO, P., ROBERT, C. P. and RYDER, R. J. (2012). Approximate Bayesian computational methods. *Stat. Comput.* **22** 1167–1180. [MR2992292](#)

- MARJORAM, P., MOLITOR, J., PLAGNOL, V. and TAVARE, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100** 15324–15328.
- MARTTINEN, P., CROUCHER, N. J., GUTMANN, M. U., CORANDER, J. and HANAGE, W. P. (2015). Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microb. Genomes* **1** e000038.
- MEEDS, E. and WELLING, M. (2014). GPS-ABC: Gaussian process surrogate approximate Bayesian computation. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*.
- NIEHUS, R., MITRI, S., FLETCHER, A. G. and FOSTER, K. R. (2015). Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nat. Commun.* **6** 8924.
- PAPAMAKARIOS, G. and MURRAY, I. (2016). Fast e-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems* 29.
- PRICE, L. F., DROVANDI, C. C., LEE, A. and NOTT, D. J. (2018). Bayesian synthetic likelihood. *J. Comput. Graph. Statist.* **27** 1–11. [MR3788296](#)
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR2514435](#)
- SHAHRIARI, B., SWERSKY, K., WANG, Z., ADAMS, R. P. and DE FREITAS, N. (2015). Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* **104**.
- SHAPIRO, B. J., FRIEDMAN, J., CORDERO, O. X., PREHEIM, S. P., TIMBERLAKE, S. C., SZABÓ, G., POLZ, M. F. and ALM, E. J. (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science* **336** 48–51.
- SISSON, S. A., FAN, Y. and TANAKA, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **104** 1760–1765. [MR2301870](#)
- SNELSON, E., RASMUSSEN, C. E. and GHAHRAMANI, Z. (2004). Warped Gaussian processes. In *Advances in Neural Information Processing Systems* 16 337–344.
- SNOEK, J., LAROCHELLE, H. and ADAMS, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems* 25 1–9.
- THOMAS, C. M. and NIELSEN, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev., Microbiol.* **3** 711–721.
- TOLVANEN, V., JYLÄNKI, P. and VEHTARI, A. (2014). Approximate inference for nonstationary heteroscedastic Gaussian process regression. In *2014 IEEE International Workshop on Machine Learning for Signal Processing* 1–24.
- 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.
- TOUCHON, M., HOEDE, C., TENAILLON, O., BARBE, V., BAERISWYL, S., BIDET, P., BINGEN, E., BONACORSI, S., BOUCHIER, C., BOUVET, O. et al. (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* **5** e1000344.
- TURNER, B. M. and SEDERBERG, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychon. Bull. Rev.* **21** 227–250.
- TURNER, B. M. and VAN ZANDT, T. (2012). A tutorial on approximate Bayesian computation. *J. Math. Psych.* **56** 69–85. [MR2909506](#)
- VANHATALO, J., RIIHIMÄKI, J., HARTIKAINEN, J., JYLÄNKI, P., TOLVANEN, V. and VEHTARI, A. (2013). GPstuff: Bayesian modeling with Gaussian processes. *J. Mach. Learn. Res.* **14** 1175–1179. [MR3063621](#)
- VEHTARI, A. and LAMPINEN, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Comput.* **14** 2439–2468.
- VEHTARI, A. and OJANEN, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Stat. Surv.* **6** 142–228. [MR3011074](#)

- WILKINSON, R. D. (2014). Accelerating ABC methods using Gaussian processes. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* 1015–1023.
- WOOD, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466** 1102–1104.

A SIMULATION-BASED FRAMEWORK FOR ASSESSING THE FEASIBILITY OF RESPONDENT-DRIVEN SAMPLING FOR ESTIMATING CHARACTERISTICS IN POPULATIONS OF LESBIAN, GAY AND BISEXUAL OLDER ADULTS

BY MARYCLARE GRIFFIN^{*,1}, KRISTA J. GILE[†],
KAREN I. FREDRICKSEN-GOLDSSEN^{*}, MARK S. HANDCOCK[‡] AND
ELENA A. EROSHEVA^{*}

University of Washington^{}, University of Massachusetts, Amherst[†]
and University of California, Los Angeles[‡]*

Respondent-driven sampling (RDS) is a method for sampling from a target population by leveraging social connections. RDS is invaluable to the study of hard-to-reach populations. However, RDS is costly and can be infeasible. RDS is infeasible when RDS point estimators have small effective sample sizes (large design effects) or when RDS interval estimators have large confidence intervals relative to estimates obtained in previous studies or poor coverage. As a result, researchers need tools to assess whether or not estimation of certain characteristics of interest for specific populations is feasible in advance. In this paper, we develop a simulation-based framework for using pilot data—in the form of a convenience sample of aggregated, egocentric data and estimates of subpopulation sizes within the target population—to assess whether or not RDS is feasible for estimating characteristics of a target population. In doing so, we assume that more is known about egos than alters in the pilot data, which is often the case with aggregated, egocentric data in practice. We build on existing methods for estimating the structure of social networks from aggregated, egocentric sample data and estimates of subpopulation sizes within the target population. We apply this framework to assess the feasibility of estimating the proportion male, proportion bisexual, proportion depressed and proportion infected with HIV/AIDS within three spatially distinct target populations of older lesbian, gay and bisexual adults using pilot data from the Caring and Aging with Pride Study and the Gallup Daily Tracking Survey. We conclude that using an RDS sample of 300 subjects is infeasible for estimating the proportion male, but feasible for estimating the proportion bisexual, proportion depressed and proportion infected with HIV/AIDS in all three target populations.

Key words and phrases. Respondent-driven sampling, hard to reach populations, social networks, aggregated egocentric sample data, network sampling.

REFERENCES

- ADMIRAAL, R. and HANDCOCK, M. S. (2016). Modeling concurrency and selective mixing in heterosexual partnership networks with applications to sexually transmitted diseases. *Ann. Appl. Stat.* **10** 2021–2046. [MR3592047](#)
- ANDRESEN, E. M., MALMGREN, J. A., CARTER, W. B. and PATRICK, D. L. (1994). Screening for depression in well older adults: Evaluation of a short form of the CES-D (Center for Epidemiologic Studies Depression Scale). *Am. J. Prev. Med.* **10** 77–84.
- BARASH, V. D., CAMERON, C. J., SPILLER, M. W. and HECKATHORN, D. D. (2016). Respondent-driven sampling—Testing assumptions: Sampling with replacement. *J. Off. Stat.* **32** 29–73.
- COHEN, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Routledge, Hillsdale, NJ.
- CORNWELL, B., LAUMANN, E. O. and SCHUMM, L. P. (2008). The social connectedness of older adults: A national profile. *Am. Sociol. Rev.* **73** 185–203.
- CRAWFORD, F. W., ARONOW, P. M., ZENG, L. and LI, J. (2018). Identification of homophily and preferential recruitment in respondent-driven sampling. *Am. J. Epidemiol.* **187** 153–160.
- EROSHEVA, E. A., KIM, H.-J., EMLET, C. and FREDRIKSEN-GOLDSSEN, K. I. (2016). Social networks of lesbian, gay, bisexual, and transgender older adults. *Research on Aging* **38** 98–123.
- FREDRIKSEN-GOLDSSEN, K. I. and MURACO, A. (2010). Aging and sexual orientation: A 25-year review of the literature. *Research on Aging* **32** 372–413.
- FREDRIKSEN-GOLDSSEN, K. I., EMLET, C. A., KIM, H.-J., MURACO, A., EROSHEVA, E. A., GOLDSSEN, J. and HOY-ELLIS, C. P. (2013). The physical and mental health of lesbian, gay male, and bisexual (LGB) older adults: The role of key health indicators and risk and protective factors. *The Gerontologist* **53** 664–675.
- FREDRIKSEN-GOLDSSEN, K. I., KIM, H.-J., SHIU, C., GOLDSSEN, J. and EMLET, C. A. (2015). Successful aging among LGBT older adults: Physical and mental health-related quality of life by age group. *The Gerontologist* **55** 154–168.
- GABOR, C. and NEPUSZ, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems* **1695** 1–9.
- GATES, G. J. (2013). LGBT Parenting in the United States. Technical report, The Williams Institute, UCLA School of Law.
- GILE, K. J. (2011). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J. Amer. Statist. Assoc.* **106** 135–146. [MR2816708](#)
- GILE, K. J. and HANDCOCK, M. S. (2010). Respondent-driven sampling: An assessment of current methodology. *Sociol. Method.* **40** 285–327.
- GILE, K. J. and HANDCOCK, M. S. (2015). Network model-assisted inference from respondent-driven sampling data. *J. Roy. Statist. Soc. Ser. A* **178** 619–639. [MR3348351](#)
- GILE, K. J., JOHNSTON, L. G. and SALGANIK, M. J. (2015). Diagnostics for respondent-driven sampling. *J. Roy. Statist. Soc. Ser. A* **178** 241–269. [MR3291770](#)
- GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 243–268. [MR2325275](#)
- GOEL, S. and SALGANIK, M. J. (2010). Assessing respondent-driven sampling. *Proc. Natl. Acad. Sci. USA* **107** 6743–6747.
- GRIFFIN, M., GILE, K. J., FREDRIKSEN-GOLDSSEN, K. I., HANDCOCK, M. S. and EROSHEVA, E. A. (2018). Supplement to “A simulation-based framework for assessing the feasibility of respondent-driven sampling for estimating characteristics in populations of lesbian, gay and bisexual older adults.” DOI:10.1214/18-AOAS1151SUPP.
- HANDCOCK, M. S., HUNTER, D. R., BUTTS, C. T., GOODREAU, S. M. and MORRIS, M. (2003). statnet: A suite of R packages for the statistical modeling of social networks. Software library.
- HANDCOCK, M. S., HUNTER, D. R., BUTTS, C. T., GOODREAU, S. M. and MORRIS, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *J. Stat. Softw.* **24** 1–29.

- HANDCOCK, M. S., HUNTER, D. R., BUTTS, C. T., GOODREAU, S. M. and MORRIS, M. (2013). ergm: Fit, simulate and analyze exponential-family models for networks. Statnet Project.
- HECKATHORN, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Soc. Probl.* **44** 174–199.
- JOHNSTON, L. G., WHITEHEAD, S., SIMIC-LAWSON, M. and KENDALL, C. (2010). Formative research to optimize respondent-driven sampling surveys among hard-to-reach populations in HIV behavioral and biological surveillance: Lessons learned from four case studies. *AIDS Care* **22** 784–792.
- KOGAN, S. M., WEJNERT, C., CHEN, Y.-F., BRODY, G. H. and SLATER, L. M. (2011). Respondent-driven sampling with hard-to-reach emerging adults: An introduction and case study with rural African americans. *Journal of Adolescent Research* **26** 30–60.
- LI, X. and ROHE, K. (2017). Central limit theorems for network driven sampling. *Electron. J. Stat.* **11** 4871–4895. [MR3733297](#)
- LOHR, S. L. (2010). *Sampling: Design and Analysis*, 2nd ed. Brooks/Cole, Cengage Learning, Boston, MA. [MR3057878](#)
- LU, X., BENGTSSON, L., BRITTON, T., CAMITZ, M., KIM, B. J., THORSON, A. and LILJEROS, F. (2012). The sensitivity of respondent-driven sampling. *J. Roy. Statist. Soc. Ser. A* **175** 191–216. [MR2873802](#)
- MALEKINEJAD, M., JOHNSTON, L. G., KENDALL, C., KERR, L. R. F. S., RIFKIN, M. R. and RUTHERFORD, G. W. (2008). Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: A systematic review. *AIDS and Behavior* **12** 105–130.
- MCCREESH, N., FROST, S. D. W., SEELEY, J., KATONGOLE, J., TARSH, M. N., NDUNGUSE, R., JICHI, F., LUNEL, N. L., MAHER, D., JOHNSTON, L. G., SONNENBERG, P., COPAS, A. J., HAYES, R. J. and WHITE, R. G. (2012). Evaluation of respondent-driven sampling. *Epidemiology* **23** 138–147.
- MERLI, M. G., MOODY, J., SMITH, J., LI, J., WEIR, S. and CHEN, X. (2015). Challenges to recruiting population representative samples of female sex workers in China using respondent driven sampling. *Social Science and Medicine* **125** 79–93.
- R CORE TEAM (2013). *R: A Language and Environment for Statistical Computing*.
- RADLOFF, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement* **1** 385–401.
- ROHE, K. (2015). Network driven sampling; a critical threshold for design effects. Preprint. Available at [arXiv:1505.05461](#).
- SALGANIK, M. J. (2006). Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health* **83** 98–112.
- SALGANIK, M. J. and HECKATHORN, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol. Method.* **34** 193–239.
- TOMAS, A. and GILE, K. J. (2011). The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *Electron. J. Stat.* **5** 899–934. [MR2831520](#)
- VERDERY, A. M., MOUW, T., BAULDRY, S. and MUCHA, P. J. (2015). Network structure and biased variance estimation in respondent driven sampling. *PLoS ONE* **10** 1–27.
- VOLZ, E. and HECKATHORN, D. D. (2008). Probability based estimation theory for respondent driven sampling. *J. Off. Stat.* **24** 79–97.
- WEJNERT, C., PHAM, H., KRISHNA, N., LE, B. and DINENNO, E. (2012). Estimating design effect and calculating sample size for respondent-driven sampling studies of injection drug users in the United States. *AIDS and Behavior* **16** 797–806.
- YE, Y. (1987). Interior algorithms for linear, quadratic, and linearly constrained non-linear programming. Ph.D. thesis, Stanford Univ.

ZEAL, M. C. (2010). Reaction to the special issue on centralizing the experiences of LGB people of color in counseling psychology. *The Counseling Psychologist* **38** 425–433.

SINGLE STAGE PREDICTION WITH EMBEDDED TOPIC MODELING OF ONLINE REVIEWS FOR MOBILE APP MANAGEMENT

BY SHAWN MANKAD^{*,1}, SHENGLI HU^{*} AND ANANDASIVAM GOPAL[†]

Cornell University^{} and University of Maryland[†]*

Mobile apps are one of the building blocks of the mobile digital economy. A differentiating feature of mobile apps to traditional enterprise software is online reviews, which are available on app marketplaces and represent a valuable source of consumer feedback on the app. We create a supervised topic modeling approach for app developers to use mobile reviews as useful sources of quality and customer feedback, thereby complementing traditional software testing. The approach is based on a constrained matrix factorization that leverages the relationship between term frequency and a given response variable in addition to co-occurrences between terms to recover topics that are both predictive of consumer sentiment and useful for understanding the underlying textual themes. The factorization is combined with ordinal regression to provide guidance from online reviews on a single app's performance as well as systematically compare different apps over time for benchmarking of features and consumer sentiment. We apply our approach using a dataset of over 100,000 mobile reviews over several years for three of the most popular online travel agent apps from the iTunes and Google Play marketplaces.

REFERENCES

- ABIResearch (2012). m-commerce growing to 24% of total e-commerce market value on back of smartphone adoption. Available at <https://www.abiresearch.com/press/m-commerce-growing-to-24-of-total-e-commerce-marke/>. Accessed: 2016-06-18.
- ABRAHAMS, A. S., FAN, W., WANG, G. A., ZHANG, Z. J. and JIAO, J. (2015). An integrated text analytic framework for product defect discovery. *Prod. Oper. Manag.* **24** 975–990.
- AGARWAL, D. and CHEN, B.-C. (2010). fLDA: Matrix factorization through latent Dirichlet allocation. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* 91–100. ACM, New York.
- ANTHES, G. (2011). Invasion of the mobile apps. *Commun. ACM* **54** 16–18.
- ARCHER, K. J. and WILLIAMS, A. A. A. (2012). L_1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Stat. Med.* **31** 1464–1474. [MR2931088](#)
- ARMSTRONG, B. G. and SLOAN, M. (1989). Ordinal regression models for epidemiologic data. *Am. J. Epidemiol.* **129** 191–204.
- ARORA, S., GE, R., HALPERN, Y., MIMNO, D., MOITRA, A., SONTAG, D., WU, Y. and ZHU, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning* 280–288.

Key words and phrases. Mobile apps, online reviews, text analysis, topic modeling, matrix factorization.

- BAVOTA, G. (2016). Mining unstructured data in software repositories: Current and future trends. In *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)* 5 1–12. IEEE, Los Alamitos, CA.
- BERTSEKAS, D. P. (1976). On the Goldstein–Levitin–Polyak gradient projection method. *IEEE Trans. Automat. Control* **21** 174–184. [MR0416017](#)
- BERTSEKAS, D. P. (1999). *Nonlinear Programming*, 2nd ed. Athena Scientific, Belmont, MA. [MR3444832](#)
- BLEI, D. M. (2012). Probabilistic topic models. *Commun. ACM* **55** 77–84.
- BLEI, D. M. and JORDAN, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1** 121–143. [MR2227367](#)
- BLEI, D. M. and LAFFERTY, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning* 113–120. ACM, New York.
- BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.
- BOYD-GRABER, J., MIMNO, D. and NEWMAN, D. (2015). Care and feeding of topic models: Problems, diagnostics, and improvements. In *Handbook of Mixed Membership Models and Their Applications*. 225–254. CRC Press, Boca Raton, FL. [MR3380033](#)
- BRODY, S. and ELHADAD, N. (2010). An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* 804–812. Association for Computational Linguistics, Stroudsburg, PA.
- BURTCH, G. and HONG, Y. (2014). What happens when word of mouth goes mobile? In *Proceedings of the International Conference on Information Systems*.
- BÜSCHKEN, J. and ALLENBY, G. M. (2016). Sentence-based text analysis for customer reviews. *Mark. Sci.* **35** 953–975.
- CAO, Q., DUAN, W. and GAN, Q. (2011). Exploring determinants of voting for the helpfulness of online user reviews: A text mining approach. *Decision Support Syst.* **50** 511–521.
- CHANG, J. (2012). *Lda: Collapsed Gibbs sampling methods for topic models*. R package version 1.3.2.
- CHEN, Z. and LURIE, N. H. (2013). Temporal contiguity and negativity bias in the impact of online word of mouth. *J. Mark. Res.* **50** 463–476.
- CHEN, N., LIN, J., HOI, S. C., XIAO, X. and ZHANG, B. (2014). AR-Miner: Mining informative reviews for developers from mobile app marketplace. In *Proceedings of the 36th International Conference on Software Engineering* 767–778. ACM, New York.
- COX, C. (1988). Multinomial regression models based on continuation ratios. *Stat. Med.* **7** 435–441.
- DEERWESTER, S. C., DUMAIS, S. T., LANDAUER, T. K., FURNAS, G. W. and HARSHMAN, R. A. (1990). Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **41** 391–407.
- DING, C., LI, T. and JORDAN, M. I. (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Mach. Intell.* **32** 45–55.
- DING, C., LI, T. and PENG, W. (2008). On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput. Statist. Data Anal.* **52** 3913–3927. [MR2432216](#)
- FEINERER, I., HORNIK, K. and MEYER, D. (2008). Text mining infrastructure in R. *J. Stat. Softw.* **25** Art. ID 5.
- FIENBERG, S. E. (2007). *The Analysis of Cross-Classified Categorical Data*. Springer, New York.
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** Art. ID 1.
- FU, B., LIN, J., LI, L., FALOUTSOS, C., HONG, J. and SADEH, N. (2013). Why people hate your app: Making sense of user feedback in a mobile app store. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1276–1284. ACM, New York.

- GALVIS CARREÑO, L. V. and WINBLADH, K. (2013). Analysis of user comments: An approach for software requirements evolution. In *Proceedings of the 2013 International Conference on Software Engineering* 582–591. IEEE Press, Los Alamitos, CA.
- GODES, D. and MAYZLIN, D. (2004). Using online conversations to study word-of-mouth communication. *Mark. Sci.* **23** 545–560.
- GRÜN, B. and HORNIK, K. (2011). topicmodels: An R package for fitting topic models. *J. Stat. Softw.* **40** Art. ID 11.
- HAN, H. J., MANKAD, S., GAVIRNENI, S. and VERMA, R. (2016). What guests really think of your hotel: Text analytics of online customer reviews. Cornell Hospitality Report, Cornell Univ., Ithaca, NY.
- HARRELL, F. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer, New York.
- HOFMANN, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 50–57. ACM, New York.
- HOLMES, D. I. (1985). The analysis of literary style—A review. *J. Roy. Statist. Soc. Ser. A* **14** 328–341.
- IACOB, C. and HARRISON, R. (2013). Retrieving and analyzing mobile apps feature requests from online reviews. In *2013 10th IEEE Working Conference on Mining Software Repositories (MSR)* 41–44. IEEE, Los Alamitos, CA.
- ICKIN, S., WAC, K., FIEDLER, M., JANOWSKI, L., HONG, J.-H. and DEY, A. K. (2012). Factors influencing quality of experience of commonly used mobile applications. *IEEE Commun. Mag.* **50** 48–56.
- JO, Y. and OH, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* 815–824. ACM, New York.
- KAN, S. H., BASILI, V. R. and SHAPIRO, L. N. (1994). Software quality: An overview from the perspective of total quality management. *IBM Syst. J.* **33** 4–19.
- KRISHNAN, M. S., KRIEBEL, C. H., KEKRE, S. and MUKHOPADHYAY, T. (2000). An empirical analysis of productivity and quality in software products. *Manage. Sci.* **46** 745–759.
- LEE, D. D. and SEUNG, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401** 788–791.
- LEE, D. D. and SEUNG, H. S. (2001). Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.* 556–562.
- LI, X. and HITT, L. M. (2008). Self-selection and information role of online product reviews. *Inf. Syst. Res.* **19** 456–474.
- LIM, S. L., BENTLEY, P. J., KANAKAM, N., ISHIKAWA, F. and HONIDEN, S. (2015). Investigating country differences in mobile app user behavior and challenges for software engineering. *IEEE Trans. Softw. Eng.* **41** 40–64.
- LIN, C.-J. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural Comput.* **19** 2756–2779. [MR2348161](#)
- LU, Y., MEI, Q. and ZHAI, C. (2011). Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. *Inf. Retr.* **14** 178–203.
- LU, B., OTT, M., CARDIE, C. and TSOU, B. K. (2011). Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW)* 81–88. IEEE, Los Alamitos, CA.
- MAALEJ, W. and NABIL, H. (2015). Bug report, feature request, or simply praise? On automatically classifying app reviews. In *2015 IEEE 23rd International Requirements Engineering Conference (RE)* 116–125. IEEE, Los Alamitos, CA.

- MANKAD, S., HU, S. and GOPAL, A. (2018). Supplement to “Single stage prediction with embedded topic modeling of online reviews for mobile app management.” DOI:10.1214/18-AOAS1152SUPP.
- MANKAD, S. and MICHAILIDIS, G. (2013). Discovery of path-important nodes using structured semi-nonnegative matrix factorization. In *2013 IEEE 5th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)* 288–291.
- MANKAD, S. and MICHAILIDIS, G. (2015). Analysis of multiview legislative networks with structured matrix factorization: Does Twitter influence translate to the real world? *Ann. Appl. Stat.* **9** 1950–1972. MR3456360
- MANKAD, S., HAN, H. J., GOH, J. and GAVIRNENI, S. (2016). Understanding online hotel reviews through automated text analysis. *Serv. Sci.* **8** 124–138.
- MCAULEY, J. and LESKOVEC, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems* 165–172. ACM, New York.
- MCAULEY, J., LESKOVEC, J. and JURAFSKY, D. (2012). Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining (ICDM)* 1020–1025. IEEE, Los Alamitos, CA.
- MCAULIFFE, J. D. and BLEI, D. M. (2008). Supervised topic models. In *Advances in Neural Information Processing Systems* 121–128.
- MCCULLAGH, P. (1980). Regression models for ordinal data. *J. Roy. Statist. Soc. Ser. B* **42** 109–142. MR0583347
- MIMNO, D., WALLACH, H. M., TALLEY, E., LEENDERS, M. and MCCALLUM, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* 262–272. Association for Computational Linguistics, Stroudsburg, PA.
- MOBILE BUSINESS INSIGHTS (2016). Mobile commerce trends: Retail in 2017, 2018 and beyond. Available at <http://mobilebusinessinsights.com/2016/12/mobile-commerce-trends-retail-in-2017-2018-and-beyond/>. Accessed: 2017-04-05.
- O’CALLAGHAN, D., GREENE, D., CATHY, J. and CUNNINGHAM, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Syst. Appl.* **42** 5645–5657.
- PANICHELLA, A., DIT, B., OLIVETO, R., DI PENTA, M., POSHYVANYK, D. and DE LUCIA, A. (2013). How to effectively use topic models for software engineering tasks? An approach based on genetic algorithms. In *Proceedings of the 2013 International Conference on Software Engineering (ICSE ’13)* 522–531. IEEE Press, Piscataway, NJ.
- PANICHELLA, S., DI SORBO, A., GUZMAN, E., VISAGGIO, C. A., CANFORA, G. and GALL, H. C. (2015). How can I improve my app? Classifying user reviews for software maintenance and evolution. In *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)* 281–290. IEEE, Los Alamitos, CA.
- PARASURAMAN, A., ZEITHAML, V. A. and BERRY, L. L. (1988). Servqual. *J. Retail.* **64** 12–40.
- PORTEOUS, I., NEWMAN, D., IHLER, A., ASUNCION, A., SMYTH, P. and WELLING, M. (2008). Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’08)* 569–577. ACM, New York.
- PRESSMAN, R. S. (2005). *Software Engineering: A Practitioner’s Approach*. Palgrave Macmillan, New York.
- ROBERTSON, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *J. Doc.* **60** 503–520.
- SALTON, G. and MICHAEL, J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.
- SERRANO, N., HERNANTES, J. and GALLARDO, G. (2013). Mobile web apps. *IEEE Softw.* **30** 22–27.

- SNYDER, B. and BARZILAY, R. (2007). Multiple aspect ranking using the good grief algorithm. In *HLT-NAACL* 300–307.
- TADDY, M. (2013). Multinomial inverse regression for text analysis. *J. Amer. Statist. Assoc.* **108** 755–770. [MR3174658](#)
- R CORE TEAM (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
- THOMAS, S. W., NAGAPPAN, M., BLOSTEIN, D. and HASSAN, A. E. (2013). The impact of classifier configuration and classifier combination on bug localization. *IEEE Trans. Softw. Eng.* **39** 1427–1443.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- TIRUNILLAI, S. and TELLIS, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *J. Mark. Res.* **51** 463–479.
- TITOV, I. and McDONALD, R. T. (2008a). A joint model of text and aspect ratings for sentiment summarization. In *ACL* **8** 308–316.
- TITOV, I. and McDONALD, R. (2008b). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web* 111–120. ACM, New York.
- WANG, C. and BLEI, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 448–456. ACM, New York.
- WANG, H., LU, Y. and ZHAI, C. (2010). Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 783–792. ACM, New York.
- WASSERMAN, A. I. (2010). Software engineering issues for mobile application development. In *Proceedings of the FSE/SDP Workshop on Future of Software Engineering Research* 397–400. ACM, New York.
- WU, Y. and ESTER, M. (2015). Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* 199–208. ACM, New York.
- XU, W., LIU, X. and GONG, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03)* 267–273. ACM, New York.

SENSITIVITY ANALYSIS FOR STRATIFIED COMPARISONS IN AN OBSERVATIONAL STUDY OF THE EFFECT OF SMOKING ON HOMOCYSTEINE LEVELS

BY PAUL R. ROSENBAUM

University of Pennsylvania

Sensitivity bounds for randomization inferences exist in several important cases, such as matched pairs with any type of outcome or binary outcomes with any type of stratification, but computationally feasible bounds for any outcome in any stratification are not currently available. For instance, with 20 strata, some large, others small, there is no currently available, computationally feasible sensitivity bound testing the null hypothesis of no treatment effect in the presence of a bias from nonrandom treatment assignment of a specific magnitude. The current paper solves the general problem; it uses an inequality formed by taking a one-step Taylor approximation from a near extreme solution, known as the separable approximation, where the concavity of the underlying function ensures that the Taylor approximation is, at worst, conservative. In practice, the separable approximation and the one-step movement away from it provide computationally feasible lower and upper bounds, thereby providing both a usable, perhaps slightly conservative statement, together with a check that the conservative statement is not unduly conservative. In every example that I have tried, the upper and lower bounds barely differ, although with some effort one can construct examples in which the separable approximation gives a P -value of 0.0499 and the Taylor approximation gives 0.0501. The new inequality holds in finite samples, so it strengthens certain existing asymptotic results, additionally simplifying the proof of those results. The method is discussed in the context of an observational study of the effects of smoking on homocysteine levels, a possible risk factor for several diseases including cardiovascular disease, thrombosis and Alzheimer's disease. This study contains two evidence factors, the comparison of smokers and nonsmokers and the comparison of smokers to one another in terms of recent nicotine exposure. A new R package, *senstrat*, implements the procedure and illustrates it with the example from the current paper.

REFERENCES

- BAZZANO, L. A., HE, J., MUNTNER, P., VUPPUTURI, S. and WHELTON, P. K. (2003). Relationship between cigarette smoking and novel risk factors for cardiovascular disease in the United States. *Ann. Intern. Med.* **138** 891–897.
- BERTSEKAS, D. P. (2009). *Convex Optimization Theory*. Athena Scientific, Nashua, NH. MR2830150

Key words and phrases. Causal inference, covariance adjustment, observational study, randomization inference, sensitivity analysis, stratification.

- BICKEL, P. J. and VAN ZWET, W. R. (1978). Asymptotic expansions for the power of distribution free tests in the two-sample problem. *Ann. Statist.* **6** 937–1004. [MR0499567](#)
- BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. [MR2061575](#)
- CENTERS FOR DISEASE CONTROL (2016). Biomonitoring summary: Cotinine. CAS No. 486-56-6. Available at https://www.cdc.gov/biomonitoring/Cotinine_BiomonitoringSummary.html, dated December 27, 2016.
- CORNFIELD, J., HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A. M., SHIMKIN, M. B. and WYNDER, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* **22** 173–203.
- EGLESTON, B. L., SCHARFSTEIN, D. O. and MACKENZIE, E. (2009). On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics* **65** 497–504. [MR2751473](#)
- FISHER, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.
- FOGARTY, C. B. and SMALL, D. S. (2016). Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *J. Amer. Statist. Assoc.* **111** 1820–1830. [MR3601738](#)
- GASTWIRTH, J. L., KRIEGER, A. M. and ROSENBAUM, P. R. (2000). Asymptotic separability in sensitivity analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 545–555. [MR1772414](#)
- GILBERT, P. B., BOSCH, R. J. and HUDGENS, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59** 531–541. [MR2004258](#)
- HANKEY, G. J. and EIKELBOOM, J. W. (1999). Homocysteine and vascular disease. *Lancet* **354** 407–413.
- HANSEN, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist. Assoc.* **99** 609–618. [MR2086387](#)
- HODGES, J. L. JR. and LEHMANN, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *Ann. Math. Stat.* **33** 482–497. [MR0156426](#)
- HOSMAN, C. A., HANSEN, B. B. and HOLLAND, P. W. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl. Stat.* **4** 849–870. [MR2758424](#)
- HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York. [MR0606374](#)
- LEHMANN, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco, CA. [MR0395032](#)
- LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. [MR2135927](#)
- LIU, W., KURAMOTO, J. and STUART, E. (2013). Sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prev. Sci.* **14** 570–580.
- MANTEL, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel–Haenszel procedure. *J. Amer. Statist. Assoc.* **58** 690–700. [MR0153079](#)
- MARITZ, J. S. (1979). A note on exact robust confidence intervals for location. *Biometrika* **66** 163–166. [MR0529161](#)
- MEHROTRA, D. V., LU, X. and LI, X. (2010). Rank-based analyses of stratified experiments: Alternatives to the van Elteren test. *Amer. Statist.* **64** 121–130. [MR2757003](#)
- NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. *Ann. Agric. Sci.* **10** 1–51. [Translated from the Polish and edited by D. M. Dąbrowska and T. P. Speed in *Statist. Sci.* **5** (1990) 465–472. [MR1092986](#)]
- PIMENTEL, S. D., SMALL, D. S. and ROSENBAUM, P. R. (2016). Constructed second control groups and attenuation of unmeasured biases. *J. Amer. Statist. Assoc.* **111** 1157–1167. [MR3561939](#)
- PURI, M. L. (1965). On the combination of independent two sample tests of a general class. *Rev. Inst. Int. Stat.* **33** 229–241. [MR0182091](#)

- ROSENBAUM, P. R. (1991). A characterization of optimal designs for observational studies. *J. Roy. Statist. Soc. Ser. B* **53** 597–610. [MR1125717](#)
- ROSENBAUM, P. R. (1995). Quantiles in nonrandom samples and observational studies. *J. Amer. Statist. Assoc.* **90** 1424–1431. [MR1379486](#)
- ROSENBAUM, P. R. (2002a). *Observational Studies*, 2nd ed. Springer, New York.
- ROSENBAUM, P. R. (2002b). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. [MR1962487](#)
- ROSENBAUM, P. R. (2014). Weighted M -statistics with superior design sensitivity in matched observational studies with multiple controls. *J. Amer. Statist. Assoc.* **109** 1145–1158. [MR3265687](#)
- ROSENBAUM, P. R. (2017a). *Observation and Experiment: An Introduction to Causal Inference*. Harvard Univ. Press, Cambridge, MA. [MR3702029](#)
- ROSENBAUM, P. R. (2017b). The general structure of evidence factors in observational studies. *Statist. Sci.* **32** 514–530. [MR3730520](#)
- ROSENBAUM, P. R. and KRIEGER, A. M. (1990). Sensitivity of two-sample permutation inferences in observational studies. *J. Amer. Statist. Assoc.* **85** 493–498.
- ROSENBAUM, P. R. and SMALL, D. S. (2017). An adaptive Mantel–Haenszel test for sensitivity analysis in observational studies. *Biometrics* **73** 422–430. [MR3665959](#)
- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.
- SATAGOPAN, J. M., OFFIT, K., FOULKES, W., ROBSON, WACHOLDER S, M. E., ENG, C. M., KARP, S. E. and BEGG, C. B. (2001). The lifetime risks of breast cancer in Ashkenazi Jewish carriers of *brca1* and *brca2* mutations. *Cancer Epidemiol. Biomark. Prev.* **10** 467–473.
- SESHADRI, S., BEISER, A., SELHUB, J., JACQUES, P. F., ROSENBERG, I. H., D’AGOSTINO, R. B., WILSON, P. W. and WOLF, P. A. (2002). Plasma homocysteine as a risk factor for dementia and Alzheimer’s disease. *N. Engl. J. Med.* **346** 476–483.
- WALD, D. S., LAW, M. and MORRIS, J. K. (2002). Homocysteine and cardiovascular disease: Evidence on causality from a meta-analysis. *Br. Med. J.* **325** 1202–1209.
- WELCH, G. N. and LOSCALZO, J. (1998). Homocysteine and atherothrombosis. *N. Engl. J. Med.* **338** 1042–1050.
- WERFEL, U., LANGEN, V., EICKHOFF, I., SCHOONBROOD, J., VAHRENHOLZ, C., BRAUKSIEPE, A., POPP, W. and NORPOTH, K. (1998). Elevated DNA single-strand breakage frequencies in lymphocytes of welders exposed to chromium and nickel. *Carcinogenesis* **19** 413–418.
- YU, B. B. and GASTWIRTH, J. L. (2005). Sensitivity analysis for trend tests: Application to the risk of radiation exposure. *Biostatistics* **6** 201–209.

VARIABLE SELECTION FOR ESTIMATING THE OPTIMAL TREATMENT REGIMES IN THE PRESENCE OF A LARGE NUMBER OF COVARIATES¹

BY BAQUN ZHANG AND MIN ZHANG²

Shanghai University of Finance and Economics and University of Michigan

Most existing methods for optimal treatment regimes, with few exceptions, focus on estimation and are not designed for variable selection with the objective of optimizing treatment decisions. In clinical trials and observational studies, often numerous baseline variables are collected and variable selection is essential for deriving reliable optimal treatment regimes. Although many variable selection methods exist, they mostly focus on selecting variables that are important for prediction (predictive variables) instead of variables that have a qualitative interaction with treatment (prescriptive variables) and hence are important for making treatment decisions. We propose a variable selection method within a general classification framework to select prescriptive variables and estimate the optimal treatment regime simultaneously. In this framework, an optimal treatment regime is equivalently defined as the one that minimizes a weighted misclassification error rate and the proposed method forward sequentially select prescriptive variables by minimizing this weighted misclassification error. A main advantage of this method is that it specifically targets selection of prescriptive variables and in the meantime is able to exploit predictive variables to improve performance. The method can be applied to both single- and multiple-decision point setting. The performance of the proposed method is evaluated by simulation studies and application to a clinical trial.

REFERENCES

- BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. [MR2216189](#)
- BARRETT, J. K., HENDERSON, R. and ROSTHØJ, C. (2014). Doubly robust estimation of optimal dynamic treatment regimes. *Stat. Biosci.* **6** 244–260.
- BIERNOT, P. and MOODIE, E. E. M. (2010). A comparison of variable selection approaches for dynamic treatment regimes. *Int. J. Biostat.* **6** Article ID 6. [MR2594878](#)
- BRINKLEY, J., TSIATIS, A. and ANSTROM, K. J. (2010). A generalized estimator of the attributable benefit of an optimal treatment regime. *Biometrics* **66** 512–522. [MR2758831](#)
- CHAKRABORTY, B., MURPHY, S. and STRECHER, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res.* **19** 317–343. [MR2757118](#)
- FAN, A., LU, W. and SONG, R. (2016). Sequential advantage selection for optimal treatment regime. *Ann. Appl. Stat.* **10** 32–53. [MR3480486](#)
- GENG, Y., ZHANG, H. H. and LU, W. (2015). On optimal treatment regimes selection for mean survival time. *Stat. Med.* **34** 1169–1184. [MR3322742](#)

Key words and phrases. C-learning, classification, high-dimensional data, misclassification error, personalized medicine, dynamic treatment regime, variable selection.

- GOLDBERG, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA.
- GUNTER, L., ZHU, J. and MURPHY, S. A. (2011). Variable selection for qualitative interactions. *Stat. Methodol.* **8** 42–55. [MR2741508](#)
- KANG, C., JANES, H. and HUANG, Y. (2014). Combining biomarkers to optimize patient treatment recommendations. *Biometrics* **70** 695–720. [MR3261788](#)
- KELLER, M. B., MCCULLOUGH, J. P., KLEIN, D. N., ARNOW, B., DUNNER, D. L., GELENBERG, A. J., MAREKOWITZ, J. C., NEMEROFF, C. B., RUSSELL, J. M., THASE, M. E., TRIVEDI, M. H. and ZAJECKA, J. (2000). A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for treatment of chronic depression. *N. Engl. J. Med.* **342** 331–336.
- LU, W., ZHANG, H. H. and ZENG, D. (2013). Variable selection for optimal treatment decision. *Stat. Methods Med. Res.* **22** 493–504. [MR3190671](#)
- MEBANE, W. R. and SEKHON, J. S. (2011). Genetic optimization using derivatives: The rgenoud package for R. *J. Stat. Softw.* **42** 1–26.
- MOODIE, E. E. M., RICHARDSON, T. S. and STEPHENS, D. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics* **63** 447–455. [MR2370803](#)
- MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 331–366. [MR1983752](#)
- PETO, R. (1982). Statistical aspects of cancer trials. In *Treatment of Cancer* (K. Halnan, ed.) 867–871. Chapman, London.
- QIAN, M. and MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210. [MR2816351](#)
- ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics* (D. Y. Halnan and P. J. Halnan, eds.). *Lect. Notes Stat.* **179** 189–326. Springer, New York. [MR2129402](#)
- ROBINS, J., ORELLANA, L. and ROTNITZKY, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Stat. Med.* **27** 4678–4721. [MR2528576](#)
- SCHULTE, P. J., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2014). Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statist. Sci.* **29** 640–661. [MR3300363](#)
- TIAN, L., ALIZADEH, A. A., GENTLES, A. J. and TIBSHIRANI, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *J. Amer. Statist. Assoc.* **109** 1517–1532. [MR3293607](#)
- WATKINS, C. J. C. H. and DAYAN, P. (1992). Q-learning. *Mach. Learn.* **8** 279–292.
- YOUNG, J. G., CAIN, L. E., ROBINS, J. M., O'REILLY, E. J. and HERNÁN, M. A. (2011). Comparative effectiveness of dynamic treatment regimes: An application of the parametric g-formula. *Stat. Biosci.* **3** 119–143.
- ZHANG, B. and ZHANG, M. (2015). C-learning: A new classification framework to estimate optimal dynamic treatment regimes. Working Paper 116, Dept. Biostatistics, Univ. Michigan, Working Paper Series.
- ZHANG, B. and ZHANG, M. (2018). Supplement to “Variable selection for estimating the optimal treatment regimes in the presence of a large number of covariates.” DOI:[10.1214/18-AOAS1154SUPP](#).
- ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2012a). A robust method for estimating optimal treatment regimes. *Biometrics* **68** 1010–1018. [MR3040007](#)
- ZHANG, B., TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LABER, E. (2012b). Estimating optimal treatment regimes from a classification perspective. *Stat.* **1** 103–114.
- ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* **100** 681–694. [MR3094445](#)

- ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** 1106–1118. [MR3010898](#)
- ZHAO, Y.-Q., ZENG, D., LABER, E. B. and KOSOROK, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.* **110** 583–598. [MR3367249](#)

MODELING HYBRID TRAITS FOR COMORBIDITY AND GENETIC STUDIES OF ALCOHOL AND NICOTINE CO-DEPENDENCE¹

BY HEPING ZHANG*, DUNGANG LIU[†], JIWEI ZHAO[‡] AND XUAN BI*

*Yale School of Public Health**, *University of Cincinnati[†]* and *State University of New York at Buffalo[‡]*

We propose a novel multivariate model for analyzing hybrid traits and identifying genetic factors for comorbid conditions. Comorbidity is a common phenomenon in mental health in which an individual suffers from multiple disorders simultaneously. For example, in the Study of Addiction: Genetics and Environment (SAGE), alcohol and nicotine addiction were recorded through multiple assessments that we refer to as hybrid traits. Statistical inference for studying the genetic basis of hybrid traits has not been well developed. Recent rank-based methods have been utilized for conducting association analyses of hybrid traits but do not inform the strength or direction of effects. To overcome this limitation, a parametric modeling framework is imperative. Although such parametric frameworks have been proposed in theory, they are neither well developed nor extensively used in practice due to their reliance on complicated likelihood functions that have high computational complexity. Many existing parametric frameworks tend to instead use pseudo-likelihoods to reduce computational burdens. Here, we develop a model fitting algorithm for the full likelihood. Our extensive simulation studies demonstrate that inference based on the full likelihood can control the type-I error rate, and gains power and improves the effect size estimation when compared with several existing methods for hybrid models. These advantages remain even if the distribution of the latent variables is misspecified. After analyzing the SAGE data, we identify three genetic variants (rs7672861, rs958331, rs879330) that are significantly associated with the comorbidity of alcohol and nicotine addiction at the chromosome-wide level. Moreover, our approach has greater power in this analysis than several existing methods for hybrid traits. Although the analysis of the SAGE data motivated us to develop the model, it can be broadly applied to analyze any hybrid responses.

REFERENCES

- ANDERSON, J. A. and PEMBERTON, J. D. (1985). The grouped continuous model for multivariate ordered categorical variables and covariate adjustment. *Biometrics* **41** 875–885. [MR0833138](#)
 BOSCARDIN, W. J., ZHANG, X. and BELIN, T. R. (2008). Modeling a mixture of ordinal and continuous repeated measures. *J. Stat. Comput. Simul.* **78** 873–886. [MR2528805](#)
 CHEN, X., CHO, K., SINGER, B. and ZHANG, H. (2011). The nuclear transcription factor PKNOX2 is a candidate gene for substance dependence in European-origin women. *PLoS ONE* **6** e16002.

Key words and phrases. Comorbidity, association, EM algorithm, latent variable, ordinal outcome.

- DE LEON, A. R. (2005). Pairwise likelihood approach to grouped continuous model and its extension. *Statist. Probab. Lett.* **75** 49–57. [MR2185609](#)
- DE LEON, A. R. and CARRIÈRE, K. C. (2007). General mixed-data model: Extension of general location and grouped continuous models. *Canad. J. Statist.* **35** 533–548. [MR2381398](#)
- DE LEON, A. R. and CARRIÈRE, K. C. (2013). *Analysis of Mixed Data: Methods and Applications*. Chapman and Hall/CRC, Boca Raton, FL.
- FERREIRA, M. A. and PURCELL, S. M. (2009). A multivariate test of association. *Bioinformatics* **25** 132–133.
- GALESLOOT, T. E., VAN STEEN, K., KIEMENEY, L. A. L. M., JANSSE, L. L. and VERMEULEN, S. H. (2014). A comparison of multivariate genome-wide association methods. *PLoS ONE* **9** e95923.
- HE, Q., AVERY, C. L. and LIN, D.-Y. (2013). A general framework for association tests with multivariate traits in large-scale genomics studies. *Genet. Epidemiol.* **37** 759–767.
- HE, J., LI, H., EDMONDSON, A. C., RADER, D. J. and LI, M. (2012). A Gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics* **13** 497–508.
- JIANG, Y., LI, N. and ZHANG, H. (2014). Identifying genetic variants for addiction via propensity score adjusted generalized Kendall's tau. *J. Amer. Statist. Assoc.* **109** 905–930. [MR3265665](#)
- KAWAKATSU, H. and LARGEY, A. G. (2009). EM algorithms for ordered probit models with endogenous regressors. *Econom. J.* **12** 164–186. [MR2500202](#)
- KIM, D. K., CHO, M. H., HERSH, C. P., LOMAS, D. A., MILLER, B. E., KONG, X., BAKKE, P., GULSVIK, A., AGUSTÍ, A., WOUTERS, E. et al. (2012). Genome-wide association analysis of blood biomarkers in chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **186** 1238–1247.
- KLEI, L., LUCA, D., DEVLIN, B. and ROEDER, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet. Epidemiol.* **32** 9–19.
- KWAK, M., ZHENG, G. and WU, C. O. (2013). Joint tests for mixed traits in genetic association studies. In *Analysis of Mixed Data* 31–41. CRC Press, Boca Raton, FL. [MR3290358](#)
- LAIRD, N. M. and LANGE, C. (2011). *The Fundamentals of Modern Statistical Genetics*. Springer, New York. [MR2762582](#)
- LANGE, C., SILVERMAN, E. K., XU, X., WEISS, S. T. and LAIRD, N. M. (2003). A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics* **4** 195–206.
- LENZ, G., DAVIS, R., NGO, V., LAM, L., GEORGE, T., WRIGHT, G., DAVE, S., ZHAO, H., XU, W., ROSENWALD, A., OTT, G., MULLER-HERMELINK, H., GASCOYNE, R., CONNORS, J., RIMSZA, L., CAMPO, E., JAFFE, E., DELABIE, J., SMELAND, E., FISHER, R., CHAN, W. and STAUDT, L. (2008). Oncogenic CARD11 mutations in human diffuse large B cell lymphoma. *Science* **319** 1676–1679.
- LI, M. D. and BURMEISTER, M. (2009). New insights into the genetics of addiction. *Nat. Rev. Genet.* **10** 225–231.
- LIU, C., RUBIN, D. B. and WU, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85** 755–770. [MR1666758](#)
- MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. [MR1243503](#)
- O'REILLY, P. F., HOGGART, C. J., POMYEN, Y., CALBOLI, F. C. F., ELLIOTT, P., JARVELIN, M.-R. and COIN, L. J. M. (2012). MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS. *PLoS ONE* **7** e34861.
- POON, W.-Y. and LEE, S. Y. (1987). Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika* **52** 409–430. [MR0914463](#)
- RUUD, P. A. (1991). Extensions of estimation methods using the EM algorithm. *J. Econometrics* **49** 305–341. [MR1129124](#)

- YANG, Q., WU, H., GUO, C.-Y. and FOX, C. S. (2010). Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genet. Epidemiol.* **34** 444–454.
- ZHANG, H. (2011). Statistical analysis in genetic studies of mental illnesses. *Statist. Sci.* **26** 116–129. [MR2866281](#)
- ZHANG, H., LIU, C.-T. and WANG, X. (2010). An association test for multiple traits based on the generalized Kendall's tau. *J. Amer. Statist. Assoc.* **105** 473–481. [MR2724840](#)
- ZHANG, H., LIU, D., ZHAO, J. and BI, X. (2018). Supplement to “Modeling hybrid traits for comorbidity and genetic studies of alcohol and nicotine co-dependence.” DOI:[10.1214/18-AOAS1156SUPP](#).
- ZHAO, J. and ZHANG, H. (2016). Modeling multiple responses via bootstrapping margins with an application to genetic association testing. *Stat. Interface* **9** 47–56. [MR3423310](#)
- ZHU, W., JIANG, Y. and ZHANG, H. (2012). Nonparametric covariate-adjusted association tests based on the generalized Kendall's tau. *J. Amer. Statist. Assoc.* **107** 1–11. [MR2949337](#)
- ZHU, W. and ZHANG, H. (2009). Why do we test multiple traits in genetic association studies? *J. Korean Statist. Soc.* **38** 1–10. [MR2656857](#)

TREE ENSEMBLES WITH RULE STRUCTURED HORSESHOE REGULARIZATION

BY MALTE NALENZ AND MATTIAS VILLANI

Linköping University

We propose a new Bayesian model for flexible nonlinear regression and classification using tree ensembles. The model is based on the RuleFit approach in Friedman and Popescu [*Ann. Appl. Stat.* **2** (2008) 916–954] where rules from decision trees and linear terms are used in a L1-regularized regression. We modify RuleFit by replacing the L1-regularization by a horseshoe prior, which is well known to give aggressive shrinkage of noise predictors while leaving the important signal essentially untouched. This is especially important when a large number of rules are used as predictors as many of them only contribute noise. Our horseshoe prior has an additional hierarchical layer that applies more shrinkage a priori to rules with a large number of splits, and to rules that are only satisfied by a few observations. The aggressive noise shrinkage of our prior also makes it possible to complement the rules from boosting in RuleFit with an additional set of trees from Random Forest, which brings a desirable diversity to the ensemble. We sample from the posterior distribution using a very efficient and easily implemented Gibbs sampler. The new model is shown to outperform state-of-the-art methods like RuleFit, BART and Random Forest on 16 datasets. The model and its interpretation is demonstrated on the well known Boston housing data, and on gene expression data for cancer classification. The posterior sampling, prediction and graphical tools for interpreting the model results are implemented in a publicly available R package.

REFERENCES

- BREIMAN, L. (1996). Stacked regressions. *Mach. Learn.* **24** 49–64.
BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.
CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2009). Handling sparsity via the horseshoe. In *AISTATS* **5** 73–80.
CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. [MR2650751](#)
CHEN, T. and GUESTRIN, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794. ACM, New York.
CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. [MR2758172](#)
COHEN, W. W. (1995). Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning (ICML'95)* 115–123.

Key words and phrases. Nonlinear regression, classification, decision trees, Bayesian, prediction, MCMC, interpretation.

- DEMBCZYŃSKI, K., KOTŁOWSKI, W. and SŁOWIŃSKI, R. (2010). ENDER: A statistical framework for boosting decision rules. *Data Min. Knowl. Discov.* **21** 52–90. [MR2720513](#)
- FREUND, Y. and SCHAPIRE, R. E. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference* **96** 148–156. Bari, Italy.
- FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. [MR1873328](#)
- FRIEDMAN, J. H. and POPESCU, B. E. (2003). Importance sampled learning ensembles. Technical report, Dept. Statistics, Stanford Univ., Stanford, CA.
- FRIEDMAN, J. H. and POPESCU, B. E. (2008). Predictive learning via rule ensembles. *Ann. Appl. Stat.* **2** 916–954. [MR2522175](#)
- FÜRNKRANZ, J. (1999). Separate-and-conquer rule learning. *Artif. Intell. Rev.* **13** 3–54.
- GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.
- GLAAB, E., GARIBALDI, J. M. and KRASNOGOR, N. (2010). Learning pathway-based decision rules to classify microarray cancer samples.
- HAHN, P. R. and CARVALHO, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *J. Amer. Statist. Assoc.* **110** 435–448. [MR3338514](#)
- KE, G., MENG, Q., FINLEY, T., WANG, T., CHEN, W., MA, W., YE, Q. and LIU, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* 3149–3157.
- LI, L. and YAO, W. (2014). Fully Bayesian logistic regression with hyper-LASSO priors for high-dimensional feature selection. *J. Stat. Comput. Simul.* **88** 2827–2851. [MR3827411](#)
- LINERO, A. R. (2018). Bayesian regression trees for high dimensional prediction and variable selection. *J. Amer. Statist. Assoc.* **113** 626–636. [MR3832214](#)
- MAKALIC, E. and SCHMIDT, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Process. Lett.* **23** 179–182.
- NALENZ, M. and VILLANI, M. (2018). Supplement to “Tree ensembles with rule structured horseshoe regularization.” DOI:[10.1214/18-AOAS1157SUPP](#).
- PIIRONEN, J. and VEHTARI, A. (2017). Comparison of Bayesian predictive methods for model selection. *Stat. Comput.* **27** 711–735. [MR3613594](#)
- 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](#)
- 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](#)
- ROKACH, L. (2010). Ensemble-based classifiers. *Artif. Intell. Rev.* **33** 1–39.
- SCHAPIRE, R. E. (1999). A brief introduction to boosting. In *IJCAI* 1401–1406.
- SINGH, D., FEBBO, P. G., ROSS, K., JACKSON, D. G., MANOLA, J., LADD, C., TAMAYO, P., RENSHAW, A. A., D’AMICO, A. V., RICHIE, J. P. et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1** 203–209.
- SLOANIM, D. K. (2002). From patterns to pathways: Gene expression data analysis comes of age. *Nat. Genet.* **32** (Supp) 502.
- SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75** 317–343.
- TERENIN, A., DONG, S. and DRAPER, D. (2016). GPU-accelerated Gibbs sampling. Preprint. Available at [arXiv:1608.04329](#).
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- VAN’T VEER, L., DAI, H., VAN DE VIJVER, M. J., HE, Y. D., HART, A. A. M., MAO, M., PETERSE, H. L., VAN DER KOY, K., MARTON, M. J., WITTEVEEN, A. T. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415** 530–536.

- WOLPERT, D. H. (1992). Stacked generalization. *Neural Netw.* **5** 241–259.
- YAP, Y., ZHANG, X., LING, M. T., WANG, X., WONG, Y. C. and DANCHIN, A. (2004). Classification between normal and tumor tissues based on the pair-wise gene expression ratio. *BMC Cancer* **4** 72.
- ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. [MR2137327](#)

MODEL TRANSFER ACROSS ADDITIVE MANUFACTURING PROCESSES VIA MEAN EFFECT EQUIVALENCE OF LURKING VARIABLES

BY ARMAN SABBAGHI¹ AND QIANG HUANG²

Purdue University and University of Southern California

Shape deviation models constitute an important component in quality control for additive manufacturing (AM) systems. However, specified models have a limited scope of application across the vast spectrum of processes in a system that are characterized by different settings of process variables, including lurking variables. We develop a new effect equivalence framework and Bayesian method that enables deviation model transfer across processes in an AM system with limited experimental runs. Model transfer is performed via inference on the equivalent effects of lurking variables in terms of an observed factor whose effect has been modeled under a previously learned process. Studies on stereolithography illustrate the ability of our framework to broaden both the scope of deviation models and the comprehensive understanding of AM systems.

REFERENCES

- BAREINBOIM, E. and PEARL, J. (2016). Causal inference and the data-fusion problem. In *Proceedings of the National Academy of Sciences* **113** 7345–7352.
- BOX, G. E. P. (1966). Use and abuse of regression. *Technometrics* **8** 625–629. [MR0203866](#)
- BUCKHOLTZ, B., RAGAI, I. and WANG, L. (2015). Cloud manufacturing: Current trends and future implementations. *J. Manuf. Sci. Eng.* **137** 040902.
- CAMPBELL, T., WILLIAMS, C., IVANOVA, O. and GARRETT, B. (2011). *Could 3D Printing Change the World? Technologies, Potential, and Implications of Additive Manufacturing*. Atlantic Council, Washington, DC.
- COOK, R. D. and CRITCHLEY, F. (2000). Identifying regression outliers and mixtures graphically. *J. Amer. Statist. Assoc.* **95** 781–794. [MR1803878](#)
- DAI, W., YANG, Q., XUE, G.-R. and YU, Y. (2007). Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*.
- DUANE, S., KENNEDY, A., PENDLETON, B. J. and ROWETH, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195** 216–222.
- GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
- GERMANY TRADE & INVEST (2014). Industrie 4.0: Smart manufacturing for the future. Online.
- GIBSON, I., ROSEN, D. W. and STUCKER, B. (2009). *Additive Manufacturing Technologies: Rapid Prototyping to Direct Digital Manufacturing*. Springer, New York.
- HUANG, Q., NOURI, H., XU, K., CHEN, Y., SOSINA, S. and DASGUPTA, T. (2014). Statistical predictive modeling and compensation of geometric deviations of 3D printed products. *J. Manuf. Sci. Eng.* **136** 061008.

Key words and phrases. 3D printing, Bayesian learning, transfer learning.

- HUANG, Q., ZHANG, J., SABBAGHI, A. and DASGUPTA, T. (2015). Optimal offline compensation of shape shrinkage for 3D printing processes. *IIE Trans. Qual. Reliab.* **47** 431–441.
- IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—For Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. [MR3309951](#)
- JIN, Y., QIN, S. and HUANG, Q. (2016). Offline predictive control of out-of-plane geometric errors for additive manufacturing. *J. Manuf. Sci. Eng.* **138** 121005.
- JOINER, B. L. (1981). Lurking variables: Some examples. *Amer. Statist.* **35** 227–233.
- LUAN, H. and HUANG, Q. (2017). Prescriptive modeling and compensation of in-plane geometric deviations for 3D printed freeform products. *IEEE Trans. Autom. Sci. Eng.* **14** 73–82.
- MENG, X.-L. (1994). Posterior predictive p -values. *Ann. Statist.* **22** 1142–1160. [MR1311969](#)
- NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng, eds.). 113–162. CRC Press, Boca Raton, FL. [MR2858447](#)
- PAN, S. J. and YANG, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22** 1345–1359.
- PARDOE, D. and STONE, P. (2010). Boosting for regression transfer. In *Proceedings of the 27th International Conference on Machine Learning*.
- PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710. [MR1380809](#)
- PEARL, J. and BAREINBOIM, E. (2014). External validity: From do-calculus to transportability across populations. *Statist. Sci.* **29** 579–595. [MR3300360](#)
- RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. [MR0760681](#)
- SABBAGHI, A., HUANG, Q. and DASGUPTA, T. (2018). Bayesian model building from small samples of disparate data for capturing in-plane deviation in additive manufacturing. *Technometrics*. DOI:10.1080/00401706.2017.1391715.
- SABBAGHI, A., DASGUPTA, T., HUANG, Q. and ZHANG, J. (2014). Inference for deformation and interference in 3D printing. *Ann. Appl. Stat.* **8** 1395–1415. [MR3271337](#)
- SHEWHART, W. A. (1931). *Economic Control of Quality of Manufacturing Product*, 1st ed. Van Nostrand Reinhold, New York.
- TONG, K., JOSHI, S. and LEHTIHET, E. A. (2008). Error compensation for fused deposition modeling (FDM) machine by correcting slice files. *Rapid Prototyping J.* **14** 4–14.
- TONG, K., LEHTIHET, E. A. and JOSHI, S. (2003). Parametric error modeling and software error compensation for rapid prototyping. *Rapid Prototyping J.* **9** 301–313.
- WANG, H. and HUANG, Q. (2006). Error cancellation modeling and its application to machining process control. *IIE Trans.* **38** 355–364.
- WANG, H. and HUANG, Q. (2007). Using error equivalence concept to automatically adjust discrete manufacturing processes for dimensional variation control. *J. Manuf. Sci. Eng.* **129** 644–652.
- WANG, H., HUANG, Q. and KATZ, R. (2005). Multi-operational machining processes modeling for sequential root cause identification and measurement reduction. *J. Manuf. Sci. Eng.* **127** 512–521.
- WU, D., ROSEN, D. W., WANG, L. and SCHAEFER, D. (2015). Cloud-based design and manufacturing: A new paradigm in digital manufacturing and design innovation. *Comput.-Aided Design* **59** 1–14.
- YATES, F. and COCHRAN, W. G. (1938). The analysis of groups of experiments. *J. Agric. Sci.* **28** 556–580.
- ZHOU, C. and CHEN, Y. (2012). Additive manufacturing based on optimized mask video projection for improved accuracy and resolution. *J. Manuf. Process.* **14** 107–118.
- ZHOU, C., CHEN, Y. and WALTZ, R. A. (2009). Optimized mask image projection for solid freeform fabrication. *J. Manuf. Sci. Eng.* **131** 061004.

SCALPEL: EXTRACTING NEURONS FROM CALCIUM IMAGING DATA

BY ASHLEY PETERSEN*, NOAH SIMON[†] AND DANIELA WITTEN[†]

*University of Minnesota** and *University of Washington[†]*

In the past few years, new technologies in the field of neuroscience have made it possible to simultaneously image activity in large populations of neurons at cellular resolution in behaving animals. In mid-2016, a huge repository of this so-called “calcium imaging” data was made publicly available. The availability of this large-scale data resource opens the door to a host of scientific questions for which new statistical methods must be developed.

In this paper we consider the first step in the analysis of calcium imaging data—namely, identifying the neurons in a calcium imaging video. We propose a dictionary learning approach for this task. First, we perform image segmentation to develop a dictionary containing a huge number of candidate neurons. Next, we refine the dictionary using clustering. Finally, we apply the dictionary to select neurons and estimate their corresponding activity over time, using a sparse group lasso optimization problem. We assess performance on simulated calcium imaging data and apply our proposal to three calcium imaging data sets.

Our proposed approach is implemented in the R package *scalpel*, which is available on CRAN.

REFERENCES

- AHRENS, M. B., ORGER, M. B., ROBSON, D. N., LI, J. M. and KELLER, P. J. (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nat. Methods* **10** 413–420.
- APTHORPE, N., RIORDAN, A., AGUILAR, R., HOMANN, J., GU, Y., TANK, D. and SEUNG, H. S. (2016). Automatic neuron detection in calcium imaging data using convolutional networks. In *Advances in Neural Information Processing Systems* 3270–3278.
- BIEN, J. and TIBSHIRANI, R. (2011). Hierarchical clustering with prototypes via minimax linkage. *J. Amer. Statist. Assoc.* **106** 1075–1084. [MR2894765](#)
- BIEN, J. and TIBSHIRANI, R. (2015). *protoclus*: Hierarchical Clustering with Prototypes. Available at <https://CRAN.R-project.org/package=protoclus>. R package version 1.5.
- CHEN, T.-W., WARDILL, T. J., SUN, Y., PULVER, S. R., RENNINGER, S. L., BAOHAN, A., SCHREITER, E. R., KERR, R. A., ORGER, M. B., JAYARAMAN, V., LOOGER, L. L., SVOBODA, K. and KIM, D. S. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499** 295–300.
- DIEGO, F. and HAMPRECHT, F. A. (2013). Learning multi-level sparse representations. In *Advances in Neural Information Processing Systems* 818–826.
- DIEGO, F. and HAMPRECHT, F. A. (2014). Sparse space–time deconvolution for calcium image analysis. In *Advances in Neural Information Processing Systems* 64–72.

Key words and phrases. Calcium imaging, cell sorting, dictionary learning, neuron identification, segmentation, clustering, sparse group lasso.

- DIEGO, F., REICHINNEK, S., BOTH, M., HAMPRECHT, F. et al. (2013). Automated identification of neuronal activity from calcium imaging by sparse dictionary learning. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on* 1058–1061. IEEE Press, New York.
- DOMBECK, D. A., KHABBAZ, A. N., COLLMAN, F., ADELMAN, T. L. and TANK, D. W. (2007). Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron* **56** 43–57.
- FRIEDRICH, J., ZHOU, P. and PANINSKI, L. (2017). Fast online deconvolution of calcium imaging data. *PLoS Comput. Biol.* **13** e1005423.
- FRIEDRICH, J., SOUDRY, D., MU, Y., FREEMAN, J., AHRES, M. and PANINSKI, L. (2015). Fast constrained non-negative matrix factorization for whole-brain calcium imaging data. In *NIPS Workshop on Statistical Methods for Understanding Neural Systems*.
- GOWER, J. C. (2006). Similarity, dissimilarity and distance, measures of. *Encyclopedia of Statistical Sciences*.
- GRIENBERGER, C. and KONNERTH, A. (2012). Imaging calcium in neurons. *Neuron* **73** 862–885.
- HAEFFELE, B., YOUNG, E. and VIDAL, R. (2014). Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* 2007–2015.
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. [MR2722294](#)
- HELMCHEN, F. and DENK, W. (2005). Deep tissue two-photon microscopy. *Nat. Methods* **2** 932–940.
- HUBER, D., GUTNISKY, D. A., PERON, S., O’CONNOR, D. H., WIEGERT, J. S., TIAN, L., OERTNER, T. G., LOOGER, L. L. and SVOBODA, K. (2012). Multiple dynamic representations in the motor cortex during sensorimotor learning. *Nature* **484** 473–478.
- JEWELL, S. and WITTEN, D. (2018). Exact spike train inference via ℓ_0 optimization. *Ann. Appl. Stat.* **12** 2457–2482.
- KO, H., HOFER, S. B., PICHLER, B., BUCHANAN, K. A., SJÖSTRÖM, P. J. and MRSIC-FLOGEL, T. D. (2011). Functional specificity of local synaptic connections in neocortical networks. *Nature* **473** 87–91.
- LOOGER, L. L. and GRIESBECK, O. (2012). Genetically encoded neural activity indicators. *Curr. Opin. Neurobiol.* **22** 18–23.
- MARUYAMA, R., MAEDA, K., MORODA, H., KATO, I., INOUE, M., MIYAKAWA, H. and AONISHI, T. (2014). Detecting cells using non-negative matrix factorization on calcium imaging data. *Neural Networks* **55** 11–19.
- MELLEN, N. M. and TUONG, C.-M. (2009). Semi-automated region of interest generation for the analysis of optically recorded neuronal activity. *Neuroimage* **47** 1331–1340.
- MISHCHENCKO, Y., VOGELSTEIN, J. T. and PANINSKI, L. (2011). A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. *Ann. Appl. Stat.* **5** 1229–1261. [MR2849773](#)
- MUKAMEL, E. A., NIMMERJAHN, A. and SCHNITZER, M. J. (2009). Automated analysis of cellular signals from large-scale calcium imaging data. *Neuron* **63** 747–760.
- OZDEN, I., LEE, H. M., SULLIVAN, M. R. and WANG, S. S.-H. (2008). Identification and clustering of event patterns from in vivo multiphoton optical recordings of neuronal ensembles. *J. Neurophysiol.* **100** 495–503.
- PACHITARIU, M., PACKER, A. M., PETTIT, N., DALGLEISH, H., HÄUSSER, M. and SAHANI, M. (2013). Extracting regions of interest from biological images with convolutional sparse block coding. In *Advances in Neural Information Processing Systems* 1745–1753.
- PANINSKI, L., PILLOW, J. and LEWIS, J. (2007). Statistical models for neural encoding, decoding, and optimal stimulus design. *Prog. Brain Res.* **165** 493–507.
- PETERSEN, A., SIMON, N. and WITTEN, D. (2018). Supplement to “SCALPEL: Extracting neurons from calcium imaging data.” DOI:[10.1214/18-AOAS1159SUPP](#).

- PNEVMATIKAKIS, E. A., SOUDRY, D., GAO, Y., MACHADO, T. A., MEREL, J., PFAU, D., REARDON, T., MU, Y., LACEFIELD, C., YANG, W. et al. (2016). Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron* **89** 285–299.
- PREVEDEL, R., YOON, Y.-G., HOFFMANN, M., PAK, N., WETZSTEIN, G., KATO, S., SCHRÖDEL, T., RASKAR, R., ZIMMER, M., BOYDEN, E. S. and VAZIRI, A. (2014). Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy. *Nat. Methods* **11** 727–730.
- ROCHEFORT, N. L., JIA, H. and KONNERTH, A. (2008). Calcium imaging in the living brain: Prospects for molecular medicine. *Trends in Molecular Medicine* **14** 389–399.
- SHEN, H. (2016). Brain-data gold mine could reveal how neurons compute. *Nature* **535** 209–210.
- SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. [MR3173712](#)
- SMITH, S. L. and HÄUSSER, M. (2010). Parallel processing of visual space by neighboring neurons in mouse visual cortex. *Nature Neuroscience* **13** 1144–1149.
- SONKA, M., HLAVAC, V. and BOYLE, R. (2014). *Image Processing, Analysis, and Machine Vision*. Cengage Learning, Boston, MA.
- SVOBODA, K. and YASUDA, R. (2006). Principles of two-photon excitation microscopy and its applications to neuroscience. *Neuron* **50** 823–839.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- VOGELSTEIN, J. T., PACKER, A. M., MACHADO, T. A., SIPPY, T., BABADI, B., YUSTE, R. and PANINSKI, L. (2010). Fast nonnegative deconvolution for spike train inference from population calcium imaging. *Journal of Neurophysiology* **104** 3691–3704.
- 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](#)
- ZHOU, P., RESENDEZ, S. L., STUBER, G. D., KASS, R. E. and PANINSKI, L. (2016). Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. Preprint. Available at [arXiv:1605.07266](#).

EXACT SPIKE TRAIN INFERENCE VIA ℓ_0 OPTIMIZATION

BY SEAN JEWELL¹ AND DANIELA WITTEN²

University of Washington

In recent years new technologies in neuroscience have made it possible to measure the activities of large numbers of neurons simultaneously in behaving animals. For each neuron a *fluorescence trace* is measured; this can be seen as a first-order approximation of the neuron’s activity over time. Determining the exact time at which a neuron spikes on the basis of its fluorescence trace is an important open problem in the field of computational neuroscience.

Recently, a convex optimization problem involving an ℓ_1 penalty was proposed for this task. In this paper we slightly modify that recent proposal by replacing the ℓ_1 penalty with an ℓ_0 penalty. In stark contrast to the conventional wisdom that ℓ_0 optimization problems are computationally intractable, we show that the resulting optimization problem can be efficiently solved for the global optimum using an extremely simple and efficient dynamic programming algorithm. Our R-language implementation of the proposed algorithm runs in a few minutes on fluorescence traces of 100,000 timesteps. Furthermore, our proposal leads to substantial improvements over the previous ℓ_1 proposal, in simulations as well as on two calcium imaging datasets.

R-language software for our proposal is available on CRAN in the package `LZeroSpikeInference`. Instructions for running this software in python can be found at <https://github.com/jewellsean/LZeroSpikeInference>.

REFERENCES

- AHRENS, M. B., ORGER, M. B., ROBSON, D. N., LI, J. M. and KELLER, P. J. (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nat. Methods* **10** 413–420.
- ALLEN INSTITUTE FOR BRAIN SCIENCE (2016). Stimulus set and response analysis. Technical report, Allen Institute, Seattle, WA.
- AUE, A. and HORVÁTH, L. (2013). Structural breaks in time series. *J. Time Series Anal.* **34** 1–16. [MR3008012](#)
- AUGER, I. E. and LAWRENCE, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.* **51** 39–54. [MR0978902](#)
- BIEN, J. and WITTEN, D. (2016). Penalized estimation in complex models. In *Handbook of Big Data*. 285–303. CRC Press, Boca Raton, FL. [MR3674823](#)
- BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. [MR2061575](#)
- 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](#)

Key words and phrases. Neuroscience, calcium imaging, changepoint detection, dynamic programming.

- BRAUN, J. V. and MÜLLER, H.-G. (1998). Statistical methods for DNA sequence segmentation. *Statist. Sci.* **13** 142–162.
- BRUNEL, N. and WANG, X.-J. (2003). What determines the frequency of fast network oscillations with irregular neural discharges? I. Synaptic dynamics and excitation-inhibition balance. *J. Neurophysiol.* **90** 415–430.
- CAVALLARI, S., PANZERI, S. and MAZZONI, A. (2016). Comparison of the dynamics of neural interactions between current-based and conductance-based integrate-and-fire recurrent networks. *Frontiers in Neural Circuits* **8**.
- CHEN, T.-W., WARDILL, T. J., SUN, Y., PULVER, S. R., RENNINGER, S. L., BAOHAN, A., SCHREITER, E. R., KERR, R. A., ORGER, M. B., JAYARAMAN, V. et al. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499** 295–300.
- DALALYAN, A. S., HEBIRI, M. and LEDERER, J. (2017). On the prediction performance of the Lasso. *Bernoulli* **23** 552–581. [MR3556784](#)
- DAVIES, P. L. and KOVAC, A. (2001). Local extremes, runs, strings and multiresolution. *Ann. Statist.* **29** 1–65. [MR1833958](#)
- DAVIS, R. A., LEE, T. C. M. and RODRIGUEZ-YAM, G. A. (2006). Structural break estimation for nonstationary time series models. *J. Amer. Statist. Assoc.* **101** 223–239. [MR2268041](#)
- DE ROOI, J. and EILERS, P. (2011). Deconvolution of pulse trains with the L_0 penalty. *Anal. Chim. Acta* **705** 218–226.
- DE ROOI, J. J., RUCKEBUSCH, C. and EILERS, P. H. C. (2014). Sparse deconvolution in one and two dimensions: Applications in endocrinology and single-molecule fluorescence imaging. *Anal. Chem.* **86** 6291–6298.
- DENEUX, T., KASZAS, A., SZALAY, G., KATONA, G., LAKNER, T., GRINVALD, A., RÓZSA, B. and VANZETTA, I. (2016). Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo. *Nat. Commun.* **7** 12190.
- DOMBECK, D. A., KHABBAZ, A. N., COLLMAN, F., ADELMAN, T. L. and TANK, D. W. (2007). Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron* **56** 43–57.
- FRIEDRICH, J. and PANINSKI, L. (2016). Fast active set methods for online spike inference from calcium imaging. In *Advances in Neural Information Processing Systems* 1984–1992.
- FRIEDRICH, J., ZHOU, P. and PANINSKI, L. (2017). Fast online deconvolution of calcium imaging data. *PLoS Comput. Biol.* **13** e1005423.
- FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42** 2243–2281. [MR3269979](#)
- GENIE PROJECT (2015). Simultaneous imaging and loose-seal cell-attached electrical recordings from neurons expressing a variety of genetically encoded calcium indicators. CRCNS.org.
- GREWE, B. F., LANGER, D., KASPER, H., KAMPA, B. M. and HELMCHEN, F. (2010). High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision. *Nat. Methods* **7** 399–405.
- HARCHAoui, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493. [MR2796565](#)
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. [MR2722294](#)
- HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations. Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. [MR3616141](#)
- HAWRYLYCZ, M., ANASTASSIOU, C., ARKHIPOV, A., BERG, J., BUICE, M., CAIN, N., GOUWENS, N. W., GRATIY, S., IYER, R., LEE, J. H. et al. (2016). Inferring cortical function in the mouse visual system through large-scale systems neuroscience. *Proc. Natl. Acad. Sci. USA* **113** 7337–7344.

- HOCKING, T. D., RIGAILL, G., FEARNEHEAD, P. and BOURQUE, G. (2017). A log-linear time algorithm for constrained changepoint detection. Preprint. Available at [ArXiv:1703.03352](#).
- HOLEKAMP, T. F., TURAGA, D. and HOLY, T. E. (2008). Fast three-dimensional fluorescence imaging of activity in neural populations by objective-coupled planar illumination microscopy. *Neuron* **57** 661–672.
- HUGELIER, S., DE ROOI, J. J., BERNEX, R., DUWÉ, S., DEVOS, O., SLIWA, M., DEDECKER, P., EILERS, P. H. and RUCKEBUSCH, C. (2016). Sparse deconvolution of high-density super-resolution images. *Sci. Rep.* **6**.
- JACKSON, B., SCARGLE, J. D., BARNES, D., ARABHI, S., ALT, A., GIOUMOUSIS, P., GWIN, E., SANGTRAKULCHAROEN, P., TAN, L. and TSAI, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Process. Lett.* **12** 105–108.
- JOHNSON, N. A. (2013). A dynamic programming algorithm for the fused lasso and L_0 -segmentation. *J. Comput. Graph. Statist.* **22** 246–260. [MR3173713](#)
- KILLICK, R., FEARNEHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. [MR3036418](#)
- LEE, C.-B. (1995). Estimating the number of change points in a sequence of independent normal random variables. *Statist. Probab. Lett.* **25** 241–248. [MR1369518](#)
- LIN, K., SHARPNACK, J., RINALDO, A. and TIBSHIRANI, R. J. (2016). Approximate recovery in changepoint problems, from ℓ_2 estimation error rates. Preprint. Available at [ArXiv:1606.06746](#).
- MAIDSTONE, R., HOCKING, T., RIGAILL, G. and FEARNEHEAD, P. (2017). On optimal multiple changepoint algorithms for large data. *Stat. Comput.* **27** 519–533. [MR3599687](#)
- MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. [MR1429931](#)
- MAZZONI, A., PANZERI, S., LOGOTHETIS, N. K. and BRUNEL, N. (2008). Encoding of naturalistic stimuli by local field potential spectra in networks of excitatory and inhibitory neurons. *PLoS Comput. Biol.* **4** e1000239, 20. [MR2476497](#)
- OLSHEN, A. B., VENKATRAMAN, E., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572.
- PNEVMATIKAKIS, E. A., MEREL, J., PAKMAN, A. and PANINSKI, L. (2013). Bayesian spike inference from calcium imaging data. In *Signals, Systems and Computers, 2013 Asilomar Conference on* 349–353. IEEE.
- PNEVMATIKAKIS, E. A., SOUDRY, D., GAO, Y., MACHADO, T. A., MEREL, J., PFAU, D., REARDON, T., MU, Y., LACEFIELD, C., YANG, W. et al. (2016). Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron* **89** 285–299.
- PREVEDEL, R., YOON, Y.-G., HOFFMANN, M., PAK, N., WETZSTEIN, G., KATO, S., SCHRÖDEL, T., RASKAR, R., ZIMMER, M., BOYDEN, E. S. et al. (2014). Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy. *Nat. Methods* **11** 727–730.
- QIAN, J. and JIA, J. (2012). On pattern recovery of the fused lasso. Preprint. Available at [ArXiv:1211.5194](#).
- RINALDO, A. (2009). Properties and refinements of the fused lasso. *Ann. Statist.* **37** 2922–2952. [MR2541451](#)
- ROJAS, C. R. and WAHLBERG, B. (2014). On change point detection using the fused lasso method. Preprint. Available at [ArXiv:1401.5408](#).
- SASAKI, T., TAKAHASHI, N., MATSUKI, N. and IKEGAYA, Y. (2008). Fast and accurate detection of action potentials from somatic calcium fluctuations. *J. Neurophysiol.* **100** 1668–1676.
- SCOTT, A. J. and KNOTT, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* **30** 507–512.
- THEIS, L., BERENS, P., FROUDARAKIS, E., REIMER, J., ROSÓN, M. R., BADEN, T., EULER, T., TOLIAS, A. S. and BETHGE, M. (2016). Benchmarking spike rate inference in population calcium imaging. *Neuron* **90** 471–482.

- 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](#)
- VAN ROSSUM, M. C. (2001). A novel spike distance. *Neural Comput.* **13** 751–763.
- VICTOR, J. D. and PURPURA, K. P. (1996). Nature and precision of temporal coding in visual cortex: A metric-space analysis. *J. Neurophysiol.* **76** 1310–1326.
- VICTOR, J. D. and PURPURA, K. P. (1997). Metric-space analysis of spike trains: Theory, algorithms and application. *Network* **8** 127–164.
- VOGELSTEIN, J. T., WATSON, B. O., PACKER, A. M., YUSTE, R., JEDYNAK, B. and PANINSKI, L. (2009). Spike inference from calcium imaging using sequential Monte Carlo methods. *Biophys. J.* **97** 636–655.
- VOGELSTEIN, J. T., PACKER, A. M., MACHADO, T. A., SIPPY, T., BABADI, B., YUSTE, R. and PANINSKI, L. (2010). Fast nonnegative deconvolution for spike train inference from population calcium imaging. *J. Neurophysiol.* **104** 3691–3704.
- VOLGUSHEV, M., ILIN, V. and STEVENSON, I. H. (2015). Identifying and tracking simulated synaptic inputs from neuronal firing: Insights from in vitro experiments. *PLoS Comput. Biol.* **11** e1004167.
- YAKSI, E. and FRIEDRICH, R. W. (2006). Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca²⁺ imaging. *Nat. Methods* **3** 377–383.
- YAO, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–189. [MR0919373](#)
- YAO, Y.-C. and AU, S. T. (1989). Least-squares estimation of a step function. *Sankhyā Ser. A* **51** 370–381. [MR1175613](#)
- ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. [MR2279469](#)

HOW OFTEN DOES THE BEST TEAM WIN? A UNIFIED APPROACH TO UNDERSTANDING RANDOMNESS IN NORTH AMERICAN SPORT

BY MICHAEL J. LOPEZ, GREGORY J. MATTHEWS AND
 BENJAMIN S. BAUMER

Skidmore College, Loyola University Chicago and Smith College

Statistical applications in sports have long centered on how to best separate signal (e.g., team talent) from random noise. However, most of this work has concentrated on a single sport, and the development of meaningful cross-sport comparisons has been impeded by the difficulty of translating luck from one sport to another. In this manuscript we develop Bayesian state-space models using betting market data that can be uniformly applied across sporting organizations to better understand the role of randomness in game outcomes. These models can be used to extract estimates of team strength, the between-season, within-season and game-to-game variability of team strengths, as well each team's home advantage. We implement our approach across a decade of play in each of the National Football League (NFL), National Hockey League (NHL), National Basketball Association (NBA) and Major League Baseball (MLB), finding that the NBA demonstrates both the largest dispersion in talent and the largest home advantage, while the NHL and MLB stand out for their relative randomness in game outcomes. We conclude by proposing new metrics for judging competitiveness across sports leagues, both within the regular season and using traditional postseason tournament formats. Although we focus on sports, we discuss a number of other situations in which our generalizable models might be usefully applied.

REFERENCES

- BAKER, R. D. and MCHALE, I. G. (2015). Time varying ratings in association football: The all-time greatest team is ... *J. Roy. Statist. Soc. Ser. A* **178** 481–492. [MR3300014](#)
- BAUMER, B. S. and MATTHEWS, G. J. (2017). *teamcolors*: Color palettes for pro sports teams. R package version 0.0.1.9001.
- BERRI, D. (2014). Noll-scully. Available at <http://wagesofwins.com/noll-scully/>. Accessed May 19, 2016.
- BERRI, D. J. and SCHMIDT, M. B. (2006). On the road with the National Basketball Association's superstar externality. *J. Sports Econ.* **7** 347–358.
- BOULIER, B. L. and STEKLER, H. O. (2003). Predicting the outcomes of National Football League games. *Int. J. Forecast.* **19** 257–270.
- BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39** 324–345. [MR0070925](#)
- BUTTREY, S. E. (2016). Beating the market betting on NHL hockey games. *J. Quant. Anal. Sports* **12** 87–98.

Key words and phrases. Sports analytics, Bayesian modeling, competitive balance, MCMC.

- CARLIN, B. P. (1996). Improved NCAA basketball tournament modeling via point spread and team strength information. *Amer. Statist.* **50** 39–43.
- CATTELAN, M., VARIN, C. and FIRTH, D. (2013). Dynamic Bradley–Terry modelling of sports tournaments. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 135–150. [MR3042315](#)
- CFP (2014). Bowl Championship Series explained. Available at http://www.collegefootballpoll.com/bcs_explained.html. Accessed May 19, 2016.
- COLQUITT, L. L., GODWIN, N. H. and CAUDILL, S. B. (2001). Testing efficiency across markets: Evidence from the NCAA basketball betting market. *J. Bus. Finance Account.* **28** 231–248.
- CRABTREE, C. (2014). NFL wary of putting Seahawks home games in prime time. Available at <http://profootballtalk.nbcsports.com/2014/04/24/nfl-wary-of-putting-seahawks-home-games-in-prime-time-due-to-recent-blowouts/>. Accessed October 19, 2016.
- CROOKER, J. R. and FENN, A. J. (2007). Sports leagues and parity when league parity generates fan enthusiasm. *J. Sports Econ.* **8** 139–164.
- DEL MORAL, P., DOUCET, A. and JASRA, A. (2006). Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 411–436. [MR2278333](#)
- DEMERS, S. (2015). Riding a probabilistic support vector machine to the Stanley Cup. *J. Quant. Anal. Sports* **11** 205–218.
- ELO, A. E. (1978). *The Rating of Chessplayers, Past and Present*. Arco Publishing, New York.
- FAHRMEIR, L. and TUTZ, G. (1994). Dynamic stochastic models for time-dependent ordered paired comparison systems. *J. Amer. Statist. Assoc.* **89** 1438–1449.
- FIRTH, D. (2017). Fair standings in soccer and other round-robin leagues. In *New England Symposium on Statistics in Sports*.
- GANDAR, J., ZUBER, R., O'BRIEN, T. and RUSSO, B. (1988). Testing rationality in the point spread betting market. *J. Finance* **43** 995–1008.
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. [MR2221284](#)
- GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. CRC Press, Boca Raton, FL. [MR3235677](#)
- GILKS, W. R. and BERZUINI, C. (2001). Following a moving target—Monte Carlo inference for dynamic Bayesian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 127–146. [MR1811995](#)
- GLICKMAN, M. E. (1995). A comprehensive guide to chess ratings. *Am. Chess J.* **3** 59–102.
- GLICKMAN, M. E. (2001). Dynamic paired comparison models with stochastic variances. *J. Appl. Stat.* **28** 673–689. [MR1862491](#)
- GLICKMAN, M. E. and STERN, H. S. (1998). A state-space model for National Football League scores. *J. Amer. Statist. Assoc.* **93** 25–35.
- GLICKMAN, M. E. and STERN, H. S. (2016). Estimating team strength in the NFL. In *Handbook of Statistical Methods and Analyses in Sports* (J. Albert, M. E. Glickman, T. B. Swartz and R. H. Koning, eds.) 5, 113–135. Chapman and Hall/CRC Press, Boca Raton, FL.
- HARVILLE, D. (1980). Predictions for National Football League games via linear-model methodology. *J. Amer. Statist. Assoc.* **75** 516–524.
- HUMPHREYS, B. R. (2002). Alternative measures of competitive balance in sports leagues. *J. Sports Econ.* **3** 133–148.
- JAMES, B., ALBERT, J. and STERN, H. S. (1993). Answering questions about baseball using statistics. *Chance* **6** 17–30.
- KNORR-HELD, L. (2000). Dynamic rating of sports teams. *J. R. Stat. Soc. Ser. D Stat.* **49** 261–276.
- KNOWLES, G., SHERONY, K. and HAUPERT, M. (1992). The demand for Major League Baseball: A test of the uncertainty of outcome hypothesis. *Am. Econ.* **36** 72–80.
- KOOPMEINERS, J. S. (2012). A comparison of the autocorrelation and variance of NFL team strengths over time using a Bayesian state-space model. *J. Quant. Anal. Sports* **8** 1–19.
- LACEY, N. J. (1990). An estimation of market efficiency in the NFL point spread betting market. *Appl. Econ.* **22** 117–129.

- LEE, Y. H. and FORT, R. (2008). Attendance and the uncertainty-of-outcome hypothesis in baseball. *Rev. Ind. Organ.* **33** 281–295.
- LEEDS, M. and VON ALLMEN, P. (2004). The economics of sports. *Bus. Sports* 361–366.
- LENTEN, L. J. (2015). Measurement of competitive balance in conference and divisional tournament design. *J. Sports Econ.* **16** 3–25.
- LOEFFELHOLZ, B., BEDNAR, E. and BAUER, K. W. (2009). Predicting NBA games using neural networks. *J. Quant. Anal. Sports* **5** Art. 7, 17. [MR2470695](#)
- LOPEZ, M. J. (2013). Inefficiencies in the national hockey league points system and the teams that take advantage. *J. Sports Econ.* **16** 410–424.
- LOPEZ, M. J. (2016). The making and comparison of draft curves. Available at <https://statsbylopez.com/2016/06/22/the-making-and-comparison-of-draft-curves/>. Accessed October 19, 2016.
- LOPEZ, M. J. and MATTHEWS, G. J. (2015). Building an NCAA men’s basketball predictive model and quantifying its success. *J. Quant. Anal. Sports* **11** 5–12.
- LOPEZ, M. J., MATTHEWS, G. J. and BAUMER, B. S. (2018). Supplement to “How often does the best team win? A unified approach to understanding randomness in North American sport.” DOI:10.1214/18-AOAS1165SUPP.
- LOPEZ, M. J. and SCHUCKERS, M. (2017). Predicting coin flips: Using resampling and hierarchical models to help untangle the NHL’s shoot-out. *J. Sports Sci.* **35** 888–897.
- MANNER, H. (2016). Modeling and forecasting the outcomes of NBA basketball games. *J. Quant. Anal. Sports* **12** 31–41.
- MASSEY, K. (1997). Statistical models applied to the rating of sports teams. Technical report, Blue-field College. Honor’s thesis.
- MATTHEWS, G. J. (2005). Improving paired comparison models for NFL point spreads by data transformation. Ph.D. thesis, Worcester Polytechnic Institute.
- MILJKOVIĆ, D., GAJIĆ, L., KOVAČEVIĆ, A. and KONJOVIĆ, Z. (2010). The use of data mining for basketball matches outcomes prediction. In *IEEE 8th International Symposium on Intelligent Systems and Informatics* 309–312. IEEE, New York.
- MOSKOWITZ, T. and WERTHEIM, L. J. (2011). *Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won*. Crown Archetype, New York, NY.
- MULLET, G. M. (1977). Simeon Poisson and the National Hockey League. *Amer. Statist.* **31** 8–12.
- NICHOLS, M. W. (2012). The impact of visiting team travel on game outcome and biases in NFL betting markets. *J. Sports Econ.* **15** 78–96.
- NOLL, R. G. (1991). Professional basketball: Economic and business perspectives. In *The Business of Professional Sports* (J. A. Mangan and P. D. Staudohar, eds.) 18–47. Univ. Illinois Press, Urbana, IL.
- OWEN, P. D. (2010). Limitations of the relative standard deviation of win percentages for measuring competitive balance in sports leagues. *Econom. Lett.* **109** 38–41.
- OWEN, A. (2011). Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA J. Manag. Math.* **22** 99–113. [MR2786228](#)
- OWEN, P. D. and KING, N. (2015). Competitive balance measures in sports leagues: The effects of variation in season length. *Econ. Inq.* **53** 731–744.
- OWEN, P. D., RYAN, M. and WEATHERSTON, C. R. (2007). Measuring competitive balance in professional team sports using the Herfindahl–Hirschman index. *Rev. Ind. Organ.* **31** 289–302.
- PAINE, N. (2013). Analyzing real home court advantage. Available at http://insider.espn.com/nba/insider/story/_/id/9014283/nba-analyzing-real-home-court-advantage-utah-jazz-denver-nuggets. Accessed October 19, 2016.
- PAUL, R. J. and WEINBACH, A. P. (2014). Market efficiency and behavioral biases in the wnba betting market. *Int. J. Financial Stud.* **2** 193–202.
- PLUMMER, M. (2016). rjags: Bayesian graphical models using MCMC. R package version 4-6.
- R CORE TEAM (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

- ROTTENBERG, S. (1956). The baseball players' labor market. *J. Polit. Econ.* **64** 242–258.
- SCULLY, G. W. (1989). *The Business of Major League Baseball*. Univ. Chicago Press, Chicago, IL.
- SOEBBING, B. P. and HUMPHREYS, B. R. (2013). Do gamblers think that teams tank? Evidence from the NBA. *Contemp. Econ. Policy* **31** 301–313.
- SPANN, M. and SKIERA, B. (2009). Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *J. Forecast.* **28** 55–72. [MR2655634](#)
- SPIEGELHALTER, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Stat. Med.* **5** 421–433.
- SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. [MR1979380](#)
- STERN, H. (1991). On the probability of winning a football game. *Amer. Statist.* **45** 179–183.
- THOMAS, A. C. (2007). Inter-arrival times of goals in ice hockey. *J. Quant. Anal. Sports* **3** Art. 5, 17. [MR2326667](#)
- TUTZ, G. and SCHAUBERGER, G. (2015). Extended ordered paired comparison models with application to football data from German Bundesliga. *AStA Adv. Stat. Anal.* **99** 209–227. [MR3328161](#)
- UTT, J. and FORT, R. (2002). Pitfalls to measuring competitive balance with Gini coefficients. *J. Sports Econ.* **3** 367–373.
- WOLFSON, J., KOOPMEINERS, J. S. and DiLERNIA, A. (2018). Who's good this year? Comparing the information content of games in the four major US sports. *J. Sports Anal.* **4** 153–163.
- YANG, T. Y. and SWARTZ, T. (2004). A two-stage Bayesian model for predicting winners in major league baseball. *J. Data Sci.* **2** 61–73.

EXTENDING BAYESIAN STRUCTURAL TIME-SERIES ESTIMATES OF CAUSAL IMPACT TO MANY-HOUSEHOLD CONSERVATION INITIATIVES¹

BY ERIC SCHMITT^{*}, CHRISTOPHER TULL[†] AND PATRICK ATWATER[‡]

Protix^{} and California Data Collaborative[†]*

Government agencies offer economic incentives to citizens for conservation actions, such as rebates for installing efficient appliances and compensation for modifications to homes. The intention of these conservation actions is frequently to reduce the consumption of a utility. Measuring the conservation impact of incentives is important for guiding policy but doing so is technically difficult. However, the methods for estimating the impact of public outreach efforts have seen substantial developments in marketing to consumers in recent years as marketers seek to substantiate the value of their services. One such method uses Bayesian Structural Time Series (BSTS) to compare a market exposed to an advertising campaign with control markets identified through a matching procedure. This paper introduces an extension of the matching/BSTS method for impact estimation to make it applicable for general conservation program impact estimation when multihousehold data is available. This is accomplished by household matching/BSTS steps to obtain conservation estimates and then aggregating the results using a meta-regression step to aggregate the findings. A case study examining the impact of rebates for household turf removal on water consumption in multiple Californian water districts is conducted to illustrate the work flow of this method.

REFERENCES

- ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2007). Synthetic control methods for comparative case studies: Estimating the effect of california’s tobacco control program. Working Paper 12831, National Bureau of Economic Research. Available at <http://www.nber.org/papers/w12831>.
- ANGRIST, J., JORDÀ, O. and KUERSTEINER, G. (2018). Semiparametric estimates of monetary policy effects: String theory revisited. *J. Bus. Econom. Statist.* **36** 371–387.
- ARELLANO, M. and BONHOMME, S. (2013). Random effects quantile regression. Technical report, CEMFI.
- BAMEZAI, A. (1995). Application of difference-in-difference techniques to the evaluation of drought-tainted water conservation programs. *Eval. Rev.* **19** 559–582.
- BRODERSEN, K. H., GALLUSSER, F., KOEHLER, J., REMY, N. and SCOTT, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *Ann. Appl. Stat.* **9** 247–274. [MR3341115](#)
- CUWCC (2015). Sustainable landscaping: Market transformation. Technical report, California Urban Water Conservation Council.
- ENDTER-WADA, J., KURTZMAN, J., KEENAN, S., KJELGREN, R. and NEALE, C. (2008). Situational waste in landscape watering: Residential and business water use in an urban Utah community. *J. Am. Water Resour. Assoc.* **44** 902–920.

Key words and phrases. Drought, water conservation, meta-analysis, time series.

- FIELDING, K. S., SPINKS, A., RUSSELL, S., MCCREA, R., STEWART, R. and GARDNER, J. (2013). An experimental test of voluntary strategies to promote urban water demand management. *J. Environ. Manag.* **114** 343–351.
- GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.
- GIORGINO, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *J. Stat. Softw.* **31** 1–24.
- HARLAN, S. L., YABIKU, S. T., LARSEN, L. and BRAZEL, A. J. (2009). Household water consumption in an arid city: Affluence, affordance, and attitudes. *Soc. Nat. Resour.* **22** 691–709.
- HENDRICKS, W. and KOENKER, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *J. Amer. Statist. Assoc.* **87** 58–68.
- HOLLIS, M. (2016). The effects of conservation messaging on retail water deliveries. *Water Practice Technol.* **11** 26–34.
- KOENKER, R. (2004). Quantile regression for longitudinal data. *J. Multivariate Anal.* **91** 74–89.
- KOENKER, R. (2005). *Quantile Regression. Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge, MA. [MR2268657](#)
- KOENKER, R. and BACHE, S. H. (2014). RQPD: Regression quantiles for panel data. Available at <https://R-Forge.R-project.org/projects/rqpd/>. R package version 0.6/r10.
- LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34** 1–41. [MR0415861](#)
- MAYER, P., LANDER, P. and GLENN, D. (2015). Outdoor water savings research initiative, Phase 1—Analysis of published research. Technical report, Alliance for Water Efficiency.
- MITCHELL, D. and CHESNUTT, T. (2013). Evaluation of east bay municipal utility district’s pilot of watersmart home water reports. Technical report, East Bay Municipal Utility District.
- MWD (2016). 2015 urban water management plan, draft march 2016. Technical report, The Metropolitan Water District of Southern California.
- OLMSTEAD, S. M., HANEMANN, W. M. and STAVINS, R. N. (2007). Water demand under alternative price structures. *J. Environ. Econ. Manag.* **54** 181–198.
- QUESNEL, K. J. and AJAMI, N. K. (2017). Changes in water consumption linked to heavy news media coverage of extreme climatic events. *Sci. Adv.* **3**(10) Article ID e1700784.
- RILEY, R. D., HIGGINS, J. P. T. and DEEKS, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ* **342** d549.
- ROMER, C. D. and ROMER, D. H. (1989). Does monetary policy matter? A new test in the spirit of Friedman and Schwartz. Working Paper 2966, National Bureau of Economic Research. Available at <http://www.nber.org/papers/w2966>.
- VIECHTBAUER, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36** 1–48.

A LOCALLY ADAPTIVE PROCESS-CONVOLUTION MODEL FOR ESTIMATING THE HEALTH IMPACT OF AIR POLLUTION

BY DUNCAN LEE¹

University of Glasgow

Most epidemiological air pollution studies focus on severe outcomes such as hospitalisations or deaths, but this underestimates the impact of air pollution by ignoring ill health treated in primary care. This paper quantifies the impact of air pollution on the rates of respiratory medication prescribed in primary care in Scotland, which is a proxy measure for the prevalence of less severe respiratory disease. A novel bivariate spatiotemporal process-convolution model is proposed, which: (i) has increased computational efficiency via a tapering function based on nearest neighbourhoods; and (ii) has locally adaptive weights that outperform traditional distance-decay kernels. The results show significant effects of particulate matter on respiratory prescription rates which are consistent with severe endpoint studies.

REFERENCES

- BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 825–848. [MR2523906](#)
- BHOPAL, R., STEINER, M., CEZARD, G., BANSAL, N., FISCHBACHER, C., SIMPSON, C., DOUGLAS, A. and SHEIKH, A. (2015). Risk of respiratory hospitalization and death, readmission and subsequent mortality: Scottish health and ethnicity linkage study. *Eur. J. Public Health* **25** 769–774.
- BLANGIARDO, M., FINAZZI, F. and CAMELETTI, M. (2016). Two-stage Bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions. *Spatial Spatio-temporal Epidemiol.* **18** 1–12.
- CENTRE FOR ENVIRONMENTAL DATA ANALYSIS (2016). HadISDH: Gridded global monthly land surface humidity data version 3.0.0.2016p.
- CENTRE FOR ENVIRONMENTAL DATA ANALYSIS (2017). Climatic Research Unit (CRU) Time-Series (TS) version 4.01 of high-resolution gridded data of month-by-month variation in climate (Jan. 1901–Dec. 2016).
- CONGDON, P. (2005). *Bayesian Models for Categorical Data*. Wiley, Chichester. [MR2191351](#)
- DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. [MR3538706](#)
- DIBBEN, C. and CLEMENS, T. (2015). Place of work and residential exposure to ambient air pollution and birth outcomes in Scotland, using geographically fine pollution climate mapping estimates. *Environ. Res.* **140** 535–541.
- DOMINICI, F., SAMET, J. and ZEGER, S. (2000). Combining evidence on air pollution and daily mortality from the 20 largest US cities: A hierarchical modelling strategy. *J. Roy. Statist. Soc. Ser. A* **163** 263–302.

Key words and phrases. Air pollution, bivariate spatiotemporal modelling, process-convolution models, respiratory medication rates.

- FURRER, R., GENTON, M. G. and NYCHKA, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15** 502–523. [MR2291261](#)
- GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. [MR2027492](#)
- HIGDON, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environ. Ecol. Stat.* **5** 173–190.
- HUANG, G., LEE, D. and SCOTT, E. (2018). Multivariate space-time modelling of multiple air pollutants and their health effects accounting for exposure uncertainty. *Stat. Med.* **37**. 1134–1148.
- LEE, D. (2018). Supplement to “A locally adaptive process-convolution model for estimating the health impact of air pollution.” DOI:[10.1214/18-AOAS1167SUPPA](#), DOI:[10.1214/18-AOAS1167SUPPB](#).
- LEE, D., MUKHOPADHYAY, S., RUSHWORTH, A. and SAHU, S. (2017). A rigorous statistical framework for spatio-temporal pollution prediction and estimation of its long-term impact on health. *Biostatistics* **18** 370–385.
- LIECHTY, J. C., LIECHTY, M. W. and MÜLLER, P. (2004). Bayesian correlation estimation. *Biometrika* **91** 1–14. [MR2050456](#)
- LOADER, A., HECTOR, D., SYKES, D., ROSE, R., TELFER, S. and GRAY, S. (2016). Air Pollution in Scotland 2015. Technical report, Ricardo Energy & Environment.
- LU, H., REILLY, C. S., BANERJEE, S. and CARLIN, B. P. (2007). Bayesian areal wombling via adjacency modeling. *Environ. Ecol. Stat.* **14** 433–452. [MR2405556](#)
- MEYER, S. and HELD, L. (2014). Power-law models for infectious disease spread. *Ann. Appl. Stat.* **8** 1612–1639. [MR3271346](#)
- RIEBLER, A., HELD, L. and RUE, H. (2012). Estimation and extrapolation of time trends in registry data—Borrowing strength from related populations. *Ann. Appl. Stat.* **6** 304–333. [MR2951539](#)
- ROYAL COLLEGE OF PHYSICIANS (2016). Every breath we take: The lifelong impact of air pollution. Available at <https://www.rcplondon.ac.uk/projects/outputs/every-breath-we-take-lifelong-impact-air-pollution>.
- WATANABE, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11** 3571–3594. [MR2756194](#)

UNCERTAINTY THROUGH THE LENSES OF A MIXED-FREQUENCY BAYESIAN PANEL MARKOV-SWITCHING MODEL

BY ROBERTO CASARIN^{*}, CLAUDIA FORONI[†], MASSIMILIANO
MARCELLINO^{‡,§} AND FRANCESCO RAVAZZOLO^{¶,||}

University Ca' Foscari of Venice^{*}, *Deutsche Bundesbank*[†], *Bocconi University,*
IGIER[‡], *C.E.P.R.*[§], *Free University of Bozen/Bolzano*[¶] and *CAMP, BI Norwegian*
Business School^{||}

We propose a Bayesian panel model for mixed frequency data, where parameters can change over time according to a Markov process. Our model allows for both structural instability and random effects. To estimate the model, we develop a Markov Chain Monte Carlo algorithm for sampling from the joint posterior distribution, and we assess its performance in simulation experiments. We use the model to study the effects of macroeconomic uncertainty and financial uncertainty on a set of variables in a multi-country context including the US, several European countries and Japan. We find that there are large differences in the effects of uncertainty in the contraction regime and the expansion regime. The use of mixed frequency data amplifies the relevance of the asymmetry. Financial uncertainty plays a more important role than macroeconomic uncertainty, and its effects are also more homogeneous across variables and countries. Disregarding either the mixed-frequency component or the Markov-switching mechanism can bring to substantially different results.

REFERENCES

- ALESSANDRI, P. and MUMTAZ, H. (2014). Financial Regimes and Uncertainty Shocks. Working Papers No. 729, Queen Mary University of London, School of Economics and Finance.
- BAKER, S. R., BLOOM, N. and DAVIS, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics* **131** 1593–1636.
- BASSETTI, F., CASARIN, R. and LEISEN, F. (2014). Beta-product dependent Pitman–Yor processes for Bayesian inference. *J. Econometrics* **180** 49–72. [MR3188911](#)
- BILLIO, M., CASARIN, R., RAVAZZOLO, F. and VAN DIJK, H. K. (2012). Combination schemes for turning point predictions. *Quarterly Review of Economics and Finance* **52** 402–412.
- BILLIO, M., CASARIN, R., COSTOLA, M. and PASQUALINI, A. (2016a). An entropy-based early warning indicator for systemic risk. *Journal of International Financial Markets, Institutions and Money* **45** 42–59.
- BILLIO, M., CASARIN, R., RAVAZZOLO, F. and VAN DIJK, H. K. (2016b). Interconnections between Eurozone and US booms and busts using a Bayesian panel Markov-switching VAR model. *J. Appl. Econometrics* **31** 1352–1370. [MR3580904](#)
- BLOOM, N. (2009). The impact of uncertainty shocks. *Econometrica* **77** 623–685. [MR2531358](#)

Key words and phrases. Dynamic panel models, econometrics, mixed-frequency, Markov-switching, Bayesian inference, Markov chain Monte Carlo.

- BLOOM, N. (2014). Fluctuations in uncertainty. *Journal of Economic Perspectives* **28** 153–76.
- CAGGIANO, G., CASTELNUOVO, E. and GROSHENNY, N. (2014). Uncertainty shocks and unemployment dynamics in U.S. recessions. *Journal of Monetary Economics* **67** 78–92.
- CANOVA, F. and CICCARELLI, M. (2004). Forecasting and turning point predictions in a Bayesian panel VAR model. *J. Econometrics* **120** 327–359. [MR2058892](#)
- CANOVA, F. and CICCARELLI, M. (2009). Estimating multicountry VAR models. *Internat. Econom. Rev.* **50** 929–959. [MR2542805](#)
- CARRIERO, A., CLARK, T. E. and MARCELLINO, M. (2018). Measuring uncertainty and its impact on the economy. *Rev. Econ. Stat.* To appear.
- CASARIN, R., FORONI, C., MARCELLINO, M. and RAVAZZOLO, F. (2018). Supplement to “Uncertainty through the lenses of a mixed-frequency Bayesian panel Markov-switching model.” DOI:10.1214/18-AOAS1168SUPPB, DOI:10.1214/18-AOAS1168SUPPC, DOI:10.1214/18-AOAS1168SUPPD, DOI:10.1214/18-AOAS1168SUPPE.
- CHIB, S. and GREENBERG, E. (1995). Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models. *J. Econometrics* **68** 339–360.
- CLARK, T. E. and RAVAZZOLO, F. (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *J. Appl. Econometrics* **30** 551–575. [MR3358636](#)
- DOVERN, J., FRITSCH, U. and SLACALEK, J. (2012). Disagreement among forecasters in G7 countries. *Rev. Econ. Stat.* **94** 1081–1096.
- FERRARA, L. and GUERIN, P. (2015). What Are The Macroeconomic Effects of High-Frequency Uncertainty Shocks? Economics Working Papers No. 2015-12, Univ. Paris West—Nanterre la Defense, EconomiX.
- FORONI, C. and MARCELLINO, M. (2014). Mixed-frequency structural models: Identification, estimation, and policy analysis. *J. Appl. Econometrics* **29** 1118–1144. [MR3293137](#)
- FORONI, C., MARCELLINO, M. and SCHUMACHER, C. (2015). Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. *J. Roy. Statist. Soc. Ser. A* **178** 57–82. [MR3291761](#)
- FORONI, C., RAVAZZOLO, F. and RIBEIRO, P. J. (2015). Forecasting commodity currencies: The role of fundamentals with short-lived predictive content. Working Paper No. 2015/14, Norges Bank.
- FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York. [MR2265601](#)
- GHYSELS, E., SANTA-CLARA, P. and VALKANOV, R. (2005). There is a risk-return trade-off after all. *J. Financ. Econ.* **76** 509–548.
- GHYSELS, E., SINKO, A. and VALKANOV, R. (2007). Midas regressions: Further results and new directions. *Econometric Rev.* **26** 53–90. [MR2339264](#)
- GUÉRIN, P. and MARCELLINO, M. (2013). Markov-switching MIDAS models. *J. Bus. Econom. Statist.* **31** 45–56. [MR3030797](#)
- JURADO, K., LUDVIGSON, S. C. and NG, S. (2015). Measuring uncertainty. *Am. Econ. Rev.* **105** 1177–1216.
- KAUFMANN, S. (2010). Dating and forecasting turning points by Bayesian clustering with dynamic structure: A suggestion with an application to Austrian data. *J. Appl. Econometrics* **25** 309–344. [MR2758637](#)
- KAUFMANN, S. (2015). *K*-state switching models with time-varying transition distributions—does loan growth signal stronger effects of variables on inflation? *J. Econometrics* **187** 82–94. [MR3347296](#)
- KHALAF, L., KICHIAN, M., SAUNDERS, C. and VOIA, M. (2013). Dynamic Panels with MIDAS Covariates: Estimation and Fit Technical Report.
- KOOP, G. and KOROBILIS, D. (2016). Forecasting With High Dimensional Panel VARs. Strathclyde Business School Working Paper.

- LIU, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Amer. Statist. Assoc.* **89** 958–966. [MR1294740](#)
- LUDVIGSON, S. C., MA, S. and NG, S. (2015). Uncertainty and Business Cycles: Exogenous Impulse or Endogenous Response? NBER Working Papers No. 21803.
- MAGNUS, J. R. and NEUDECKER, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, Chichester. [MR1698873](#)
- MIN, C. and ZELLNER, A. (1993). Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *J. Econometrics* **56** 89–118.
- MUMTAZ, H. and THEODORIDIS, K. (2018). The changing transmission of uncertainty shocks in the US. *J. Bus. Econom. Statist.* **36** 239–252. [MR3790211](#)
- PETTENUZZO, D., TIMMERMANN, A. and VALKANOV, R. (2016). A MIDAS approach to modeling first and second moment dynamics. *J. Econometrics* **193** 315–334. [MR3515896](#)
- ROBERT, C. P. and CASELLA, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York. [MR1707311](#)
- ROBERTS, G. O. and SAHU, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. Roy. Statist. Soc. Ser. B* **59** 291–317. [MR1440584](#)
- RODRIGUEZ, A. and PUGGIONI, G. (2010). Mixed frequency models: Bayesian approaches to estimation and prediction. *Int. J. Forecast.* **26** 293–311.
- ROSSI, B. and SEKHPOSYAN, T. (2015). Macroeconomic uncertainty indices based on nowcast and forecast error distributions. *Am. Econ. Rev.* **105** 650–655.
- SCOTTI, C. (2016). Surprise and uncertainty indexes: Real-time aggregation of real-activity macro-surprises. *Journal of Monetary Economics* **82** 1–19.

DISENTANGLING AND ASSESSING UNCERTAINTIES IN MULTIPERIOD CORPORATE DEFAULT RISK PREDICTIONS

BY MIAO YUAN*, CHENG YONG TANG^{†,1}, YILI HONG*,² AND JIAN YANG^{‡,3}

*Virginia Tech**, *Temple University[†]* and *University of Colorado Denver[‡]*

Measuring the corporate default risk is broadly important in economics and finance. Quantitative methods have been developed to predictively assess future corporate default probabilities. However, as a more difficult yet crucial problem, evaluating the uncertainties associated with the default predictions remains little explored. In this paper, we attempt to fill this blank by developing a procedure for quantifying the level of associated uncertainties upon carefully disentangling multiple contributing sources. Our framework effectively incorporates broad information from historical default data, corporations' financial records, and macroeconomic conditions by (a) characterizing the default mechanism, and (b) capturing the future dynamics of various features contributing to the default mechanism. Our procedure overcomes the major challenges in this large scale statistical inference problem and makes it practically feasible by using parsimonious models, innovative methods, and modern computational facilities. By predicting the marketwide total number of defaults and assessing the associated uncertainties, our method can also be applied for evaluating the aggregated market credit risk level. Upon analyzing a US market data set, we demonstrate that the level of uncertainties associated with default risk assessments is indeed substantial. More informatively, we also find that the level of uncertainties associated with the default risk predictions is correlated with the level of default risks, indicating potential for new scopes in practical applications including improving the accuracy of default risk assessments.

REFERENCES

- ALTMAN, E. I. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. *J. Finance* **23** 589–609.
- ALTMAN, E., RESTI, A. and SIRONI, A. (2004). Default recovery rates in credit risk modelling: A review of the literature and empirical evidence. *Economic Notes* **33** 183–208.
- BAI, J. and NG, S. (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics* **3** 89–163.
- BAI, J. and WANG, P. (2015). Identification and Bayesian estimation of dynamic factor models. *J. Bus. Econom. Statist.* **33** 221–240. [MR3337059](#)
- BAÑBURA, M. and MODUGNO, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *J. Appl. Econometrics* **29** 133–160. [MR3233736](#)
- BEAVER, W. H. (1966). Financial ratios as predictors of failure. *J. Acc. Res.* **4** 71.
- BEAVER, W. H. (1968). Market prices, financial ratios, and the prediction of failure. *J. Acc. Res.* **6** 179.

Key words and phrases. Competing risks, corporate default probability, EM algorithm, dynamic factor model, high-dimensional time series, prediction interval.

- BEAVER, W. H., CORREIA, M. and MCNICHOLS, M. F. (2012). Do differences in financial reporting attributes impair the predictive ability of financial ratios for bankruptcy? *Rev. Acc. Stud.* **17** 969–1010.
- BHARATH, S. T. and SHUMWAY, T. (2008). Forecasting default with the merton distance to default model. *Rev. Financ. Stud.* **21** 1339–1369.
- BÖHNING, D., DIETZ, E., SCHAUB, R., SCHLATTMANN, P. and LINDSAY, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Ann. Inst. Statist. Math.* **46** 373–388.
- BÄRUNING, F. and KOOPMAN, S. J. (2014). Forecasting macroeconomic variables using collapsed dynamic factor analysis. *Int. J. Forecast.* **30** 572–584.
- CAMPBELL, J. Y., HILSCHER, J. and SZILAGYI, J. (2008). In search of distress risk. *J. Finance* **63** 2899–2939.
- CHAVA, S. and JARROW, R. A. (2004). Bankruptcy prediction with industry effects. *Rev. Finance* **8** 537–569.
- DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **13** 253–263.
- DING, A. A., TIAN, S., YU, Y. and GUO, H. (2012). A class of discrete transformation survival models with application to default probability prediction. *J. Amer. Statist. Assoc.* **107** 990–1003. [MR3010885](#)
- DUAN, J.-C., SUN, J. and WANG, T. (2012). Multiperiod corporate default prediction—A forward intensity approach. *J. Econometrics* **170** 191–209. [MR2955949](#)
- DUFFIE, D. (2011). *Measuring Corporate Default Risk*. Oxford Univ. Press, Oxford.
- DUFFIE, D. and LANDO, D. (2001). Term structures of credit spreads with incomplete accounting information. *Econometrica* **69** 633–664. [MR1828538](#)
- DUFFIE, D., SAITA, L. and WANG, K. (2007). Multi-period corporate default prediction with stochastic covariates. *J. Financ. Econ.* **83** 635–665.
- DUFFIE, D., ECKNER, A., HOREL, G. and SAITA, L. (2009). Frailty correlated default. *J. Finance* **64** 2089–2123.
- DURBIN, J. and KOOPMAN, S. J. (2012). *Time Series Analysis by State Space Methods*, 2nd ed. *Oxford Statistical Science Series* **38**. Oxford Univ. Press, Oxford. [MR3014996](#)
- GIESECKE, K. and KIM, B. (2011). Systemic risk: What defaults are telling us. *Manage. Sci.* **57** 1387–1405.
- GIESECKE, K., LONGSTAFF, F., SCHAFER, S. and STREBULAIEV, I. (2011). Corporate bond default risk: A 150-year perspective. *J. Financ. Econ.* **102** 233–250.
- HILLEGEIST, S. A., KEATING, E. K., CRAM, D. P. and LUNDSTEDT, K. G. (2004). Assessing the probability of bankruptcy. *Rev. Acc. Stud.* **9** 5–34.
- HONG, Y. (2013). On computing the distribution function for the Poisson binomial distribution. *Comput. Statist. Data Anal.* **59** 41–51. [MR3000040](#)
- HONG, Y., MEEKER, W. Q. and MCCALLEY, J. D. (2009). Prediction of remaining life of power transformers based on left truncated and right censored lifetime data. *Ann. Appl. Stat.* **3** 857–879. [MR2750685](#)
- HOSMER, D. W., LEMESHOW, S. and STURDIVANT, R. X. (2013). *Applied Logistic Regression*. John Wiley and Sons Ltd.
- KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. [MR1924807](#)
- KOOPMAN, S. J. and LUCAS, A. (2005). Business and default cycles for credit risk. *J. Appl. Econometrics* **20** 311–323. [MR2221901](#)
- KOOPMAN, S. J. and LUCAS, A. (2008). A non-Gaussian panel time series model for estimating and decomposing default risk. *J. Bus. Econom. Statist.* **26** 510–525. [MR2459348](#)
- KOOPMAN, S. J., LUCAS, A. and MONTEIRO, A. (2008). The multi-state latent factor intensity model for credit rating transitions. *J. Econometrics* **142** 399–424. [MR2408742](#)

- KOOPMAN, S. J., LUCAS, A. and SCHWAAB, B. (2011). Modeling frailty-correlated defaults using many macroeconomic covariates. *J. Econometrics* **162** 312–325. [MR2795620](#)
- KOOPMAN, S. J., LUCAS, A. and SCHWAAB, B. (2012). Dynamic factor models with macro, frailty, and industry effects for U.S. default counts: The credit crisis of 2008. *J. Bus. Econom. Statist.* **30** 521–532. [MR2990495](#)
- LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40** 694–726. [MR2933663](#)
- LAWLESS, J. F. and FREDETTE, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika* **92** 529–542. [MR2202644](#)
- MEEKER, W. Q. and ESCOBAR, L. A. (1998). *Statistical Methods for Reliability Data*. Wiley, New York.
- MERTON, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *J. Finance* **29** 449–470.
- OHLSON, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *J. Acc. Res.* **18** 109.
- PAN, J. and YAO, Q. (2008). Modelling multiple time series via common factors. *Biometrika* **95** 365–379. [MR2521589](#)
- PENG, X. and KOU, S. (2009). Default clustering and valuation of collateralized debt obligations. Working paper.
- SCHWAAB, B., KOOPMAN, S. J. and LUCAS, A. (2017). Global credit risk: World, country and industry factors. *J. Appl. Econometrics* **32** 296–317. [MR3631944](#)
- SHUMWAY, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *J. Bus.* **74** 101–124.
- STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179. [MR1951271](#)
- TSAY, R. S. (2010). *Analysis of Financial Time Series*, 3rd ed. Wiley, Hoboken, NJ. [MR2778591](#)
- TSAY, R. S. (2014). *Multivariate Time Series Analysis: With R and Financial Applications*. Wiley, Hoboken, NJ. [MR3236787](#)
- VOLKOVA, A. Y. (1996). A refinement of the central limit theorem for sums of independent random indicators. *Theory Probab. Appl.* **40** 791–794. [MR1405154](#)
- YUAN, M., TANG, C. Y., HONG, Y. and YANG, J. (2018). Supplement to “Disentangling and assessing uncertainties in multiperiod corporate default risk predictions.” DOI:[10.1214/18-AOAS1170SUPP](#).
- ZMIJEWSKI, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *J. Acc. Res.* **22** 59.

ON THE EVOLUTION OF THE UNITED KINGDOM PRICE DISTRIBUTIONS¹

BY BA CHU^{*}, KIM HUYNH[†], DAVID JACHO-CHÁVEZ[‡] AND
OLEKSIY KRYVTSOV[†]

Carleton University^{}, Bank of Canada[†] and Emory University[‡]*

We propose a functional principal components method that accounts for stratified random sample weighting and time dependence in the observations to understand the evolution of distributions of monthly micro-level consumer prices for the United Kingdom (UK). We apply the method to publicly available monthly data on individual-good prices collected in retail stores by the UK Office for National Statistics for the construction of the UK Consumer Price Index from March 1996 to September 2015. In addition, we conduct Monte Carlo simulations to demonstrate the effectiveness of our methodology. Our method allows us to visualize the dynamics of the price distribution and uncovers interesting patterns during the sample period. Further, we demonstrate the efficacy of our methodology with an out-of-sample forecasting algorithm which exploits the time dependence of distributions. Our out-of-sample forecasts compares favorably with the random walk forecast.

REFERENCES

- ALVAREZ, F., LE BIHAN, H. and LIPPI, F. (2016). The real effects of monetary shocks in sticky price models: A sufficient statistic approach. *Am. Econ. Rev.* **106** 2817–51. DOI:[10.1257/aer.20140500](https://doi.org/10.1257/aer.20140500).
- BALL, L. and MANKIW, N. G. (1995). Relative-price changes as aggregate supply shocks. *Q. J. Econ.* **110** 161. DOI:[10.2307/2118514](https://doi.org/10.2307/2118514).
- BELLHOUSE, D. R. and STAFFORD, J. E. (1999). Density estimation from complex surveys. *Statist. Sinica* **9** 407–424. [MR1707847](https://doi.org/10.1214/07-SS11)
- BENKO, M., HÄRDLE, W. and KNEIP, A. (2009). Common functional principal components. *Ann. Statist.* **37** 1–34. [MR2488343](https://doi.org/10.1214/08-AN.123)
- BERGER, D. and VAVRA, J. (2018). Dynamics of the U.S. price distribution. *Eur. Econ. Rev.* **103** 60–82.
- BILS, M. and KLENOW, P. J. (2004). Some evidence on the importance of price stickiness. *J. Polit. Econ.* **112** 947–987.
- BUSKIRK, T. D. (1998). Nonparametric density estimation using complex survey data. In *Proceedings of the Survey Research Methods Section* 799–801. Amer. Statist. Assoc., Alexandria, VA.
- BUSKIRK, T. D. (1999). Using nonparametric methods for density estimation with complex survey data. Ph.D. dissertation, Arizona State Univ. [MR2699315](https://doi.org/10.1214/07-SS11)
- BUSKIRK, T. D. and LOHR, S. L. (2005). Asymptotic properties of kernel density estimation with complex survey data. *J. Statist. Plann. Inference* **128** 165–190. [MR2110183](https://doi.org/10.1214/05-SP.1)
- CANTY, A. J. and DAVISON, A. C. (1999). Resampling-based variance estimation for labour force surveys. *Statistician* **48** 379–391.

Key words and phrases. Consumer price distributions, nonparametric methods, Functional Principal Component Analysis (FPCA), stratified random sampling, strong mixing.

- CASTRO CAMILO, D. and DE CARVALHO, M. (2017). Spectral density regression for bivariate extremes. *Stoch. Environ. Res. Risk Assess.* **31** 1603–1613.
- CHEN, H. A., LEVY, D., RAY, S. and BERGEN, M. (2008). Asymmetric price adjustment in the small. *J. Econ. Perspect.* **55** 728–737.
- CHU, B. M., HUYNH, K. P., JACHO-CHÁVEZ, D. T. and KRYVTSOV, O. (2018). Supplement to “On the evolution of the United Kingdom price distributions.” DOI:[10.1214/18-AOAS1172SUPP](https://doi.org/10.1214/18-AOAS1172SUPP).
- CLEWS, G., SANDERSON, R. and RALPH, J. (2014). Calculating a retrospective superlative Consumer Prices Index for the UK. Technical report, Office for National Statistics.
- DEBELLE, G. and LAMONT, O. (1997). Relative price variability and inflation: Evidence from U.S. cities. *J. Polit. Econ.* **105** 132–152.
- DHYNE, E., ALVAREZ, L., LE BIHAN, H., VERONESE, G., DIAS, D., HOFFMANN, J., JONKER, N., LUNNEMANN, P., RUMLER, F. and VILMUNEN, J. (2006). Price changes in the euro area and the United States: Some facts from individual consumer price data. *J. Econ. Perspect.* **20** 171–192.
- DIEWERT, E. W. (2012). Consumer price statistics in the UK. Report for the UK Office for National Statistics.
- DUIN, R. P. W. (1976). On the choice of smoothing parameters for parzen estimators of probability density functions. *IEEE Trans. Comput.* **25** 1175–1179.
- GAGNON, E. (2009). Price setting during low and high inflation: Evidence from Mexico. *Q. J. Econ.* **124** 1221–1263.
- VAVRA, J. S. (2011). Inflation dynamics and time-varying volatility: new evidence and an Ss interpretation. *Q. J. Econ.* **129** 215–258.
- HUYNH, K. P. and JACHO-CHÁVEZ, D. T. (2010). Firm size distributions through the lens of functional principal components analysis. *J. Appl. Econometrics* **25** 1211–1214. [MR2757237](#)
- HUYNH, K. P., JACHO-CHÁVEZ, D. T., PETRUNIA, R. J. and VOIA, M. (2011). Functional principal component analysis of density families with categorical and continuous data on Canadian entrant manufacturing firms. *J. Amer. Statist. Assoc.* **106** 858–878. [MR2894743](#)
- HUYNH, K. P., JACHO-CHÁVEZ, D. T., KRYVTSOV, O., SHEPOTYLO, O. and VAKHITOV, V. (2016). The evolution of firm-level distributions for Ukrainian manufacturing firms. *J. Comp. Econ.* **44** 148–162.
- HYNDMAN, R. J., BASHTANNYK, D. M. and GRUNWALD, G. K. (1996). Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.* **5** 315–336. [MR1422114](#)
- HYNDMAN, R. J. and KHANDAKAR, Y. (2008). Automatic time series forecasting: The forecast package for R. *J. Stat. Softw.* **27** 1–22.
- KAPLAN, G. and MENZIO, G. (2015). The morphology of price dispersion. *Internat. Econom. Rev.* **56** 1165–1206.
- KLENOW, P. J. and KRYVTSOV, O. (2008). State-dependent or time-dependent pricing: Does it matter for recent U.S. inflation? *Q. J. Econ.* **123** 863–904.
- KLENOW, P. J. and MALIN, B. (2010). Microeconomic evidence on price-setting. In *Handbook of Monetary Economics* (B. M. Friedman and M. Woodford, eds.) **3** 231–284. Elsevier, Amsterdam.
- KNEIP, A. and UTIKAL, K. J. (2001). Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc.* **96** 519–542. [MR1946423](#)
- KRYVTSOV, O. (2016). Is there a quality bias in the Canadian CPI? Evidence from microdata. *Can. J. Econ.* **49** 1401–1424.
- KRYVTSOV, O. and VINCENT, N. (2017). The cyclicalities of sales and aggregate price flexibility. Mimeo.
- LACH, S. and TSIDDON, D. (1992). The behavior of prices and inflation: An empirical analysis of disaggregated price data. *J. Polit. Econ.* **100** 349–389.
- MIDRIGAN, V. (2011). Menu costs, multiproduct firms, and aggregate fluctuations. *Econometrica* **79** 1139–1180. [MR2857882](#)

- NAKAMURA, E. and STEINSSON, J. (2008). Five facts about prices: A reevaluation of menu cost models. *Q. J. Econ.* **123** 1415–1464.
- NAKAMURA, E. and STEINSSON, J. (2013). Price rigidity: Microeconomic evidence and macroeconomic implications. *Ann. Rev. Econ.* **213** 133–163.
- ONS (2014). Consumer Price Indices—Technical Manual. The Office of National Statistics, Prices Division, Newport, UK.
- PETERSON, B. and SHI, S. (2004). Money, price dispersion and welfare. *Econom. Theory* **24** 907–932. [MR2092968](#)
- PIESSENS, R., DE DONCKER-KAPENGA, E., ÜBERHUBER, C. W. and KAHANER, D. K. (1983). *QUADPACK: A Subroutine Package for Automatic Integration*. Springer Series in Computational Mathematics **1**. Springer, Berlin. [MR0712135](#)
- ROSENBLATT, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. USA* **42** 43–47. [MR0074711](#)
- SHEREMIROV, V. (2016). Price dispersion and inflation: New facts and theoretical implications. Mimeo.
- SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London. [MR0848134](#)
- THORARINSDOTTIR, T. L., GNEITING, T. and GISSIBL, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA J. Uncertain. Quantificat.* **1** 522–534. [MR3283896](#)
- TRAN, L. T. (1990). Kernel density estimation on random fields. *J. Multivariate Anal.* **34** 37–53. [MR1062546](#)

INSTRUMENTAL VARIABLE ANALYSIS WITH CENSORED DATA IN THE PRESENCE OF MANY WEAK INSTRUMENTS: APPLICATION TO THE EFFECT OF BEING SENTENCED TO PRISON ON TIME TO EMPLOYMENT

BY ASHKAN ERTEFAIE^{*}, ANH NGUYEN[†], DAVID J. HARDING[‡],
JEFFREY D. MORENOFF^{†,1} AND WEI YANG[§]

University of Rochester^{*}, *University of Michigan*[†], *University of California,*
Berkeley[‡] and *University of Pennsylvania*[§]

This article discusses an instrumental variable approach for analyzing censored data that includes many instruments that are weakly associated with the endogenous variable. We study the effect of imprisonment on time to employment using an administrative data on all individuals sentenced for felony in Michigan in the years 2003–2006. Despite the large body of research on the effect of prison on employment, this is still a controversial topic, especially since some of the studies could have been affected by unmeasured confounding. We take advantage of a natural experiment based on the random assignment of judges to felony cases and construct a vector of instruments based on judges' ID that can avoid the confounding bias. However, some of the constructed instruments are weakly associated with the sentence type, that is, the endogenous variable, which can potentially lead to misleading results. Using a dimension reduction technique, we propose a novel semi-parametric estimation procedure in a survival context that is robust to the presence of many weak instruments. Specifically, we construct a test statistic based on the structural failure time model and provide inference by inverting the testing procedure. Under some assumptions, the optimal choice of the test statistic has also been derived. Analyses show a significant negative impact of imprisonment on time to employment which is consistent with some of the previous results. Our simulation studies highlight the importance of accounting for weak instruments in the analyses in terms of both bias and inflated type-I error rates.

REFERENCES

- ABADIE, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *J. Econometrics* **113** 231–263. [MR1960380](#)
- ALEXANDER, M. (2012). *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press.
- ANDERSON, T. W., KUNITOMO, N. and MATSUSHITA, Y. (2010). On the asymptotic optimality of the LIML estimator with possibly many instruments. *J. Econometrics* **157** 191–204. [MR2661594](#)
- ANDERSON, T. W. and RUBIN, H. (1950). The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Stat.* **21** 570–582. [MR0039188](#)

Key words and phrases. Instrumental variables, survival time, test statistics, two-stage least squares, unmeasured confounders.

- ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.
- APEL, R. and SWEETEN, G. (2010). The impact of incarceration on employment during the transition to adulthood. *Soc. Probl.* **57** 448–479.
- AUSTIN, J. and IRWIN, J. (2012). *It's About Time: America's Imprisonment Binge*. Cengage Learning, Boston, MA.
- BAIOCCHI, M., SMALL, D. S., LORCH, S. and ROSENBAUM, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Amer. Statist. Assoc.* **105** 1285–1296. [MR2796550](#)
- BEKKER, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* **62** 657–681. [MR1281697](#)
- BELLONI, A., CHEN, D., CHERNOZHUKOV, V. and HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80** 2369–2429. [MR3001131](#)
- BOSCO, J. L. F., SILLIMAN, R. A., THWIN, S. S., GEIGER, A. M., BUIST, D. S. M., PROUT, M. N., YOOD, M. U., HAQUE, R., WEI, F. and LASH, T. L. (2010). A most stubborn bias: No adjustment method fully resolves confounding by indication in observational studies. *J. Clin. Epidemiol.* **63** 64–74.
- BOUND, J., JAEGER, D. A. and BAKER, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Amer. Statist. Assoc.* **90** 443–450.
- BROOKHART, M. A. and SCHNEEWEISS, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *Int. J. Biostat.* **3** Art. 14, 25. [MR2383610](#)
- BROOKHART, M. A., WANG, P., SOLOMON, D. H. and SCHNEEWEISS, S. (2006). Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* **17** 268.
- BUSE, A. (1992). The bias of instrumental variable estimators. *Econometrica* **60** 173–180. [MR1161549](#)
- CHAMBERLAIN, G. and IMBENS, G. (2004). Random effects estimators with many instrumental variables. *Econometrica* **72** 295–306. [MR2031020](#)
- CHAN, K. C. G. (2016). Instrumental variable additive hazards models with exposure-dependent censoring. *Biometrics*.
- CHAO, J. C. and SWANSON, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica* **73** 1673–1692. [MR2156676](#)
- CHAO, J. C., HAUSMAN, J. A., NEWEY, W. K., SWANSON, N. R. and WOUTERSEN, T. (2014). Testing overidentifying restrictions with many instruments and heteroskedasticity. *J. Econometrics* **178** 15–21. [MR3137889](#)
- CHOI, I. and PHILLIPS, P. C. B. (1992). Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified structural equations. *J. Econometrics* **51** 113–150. [MR1151956](#)
- CONLEY, T. G., HANSEN, C. B. and ROSSI, P. E. (2012). Plausibly exogenous. *Rev. Econ. Stat.* **94** 260–272.
- CURRIE, E. (2013). *Crime and Punishment in America*. Macmillan, New York.
- DAVIDSON, R. and MACKINNON, J. G. (2014). Confidence sets based on inverting Anderson–Rubin tests. *Econom. J.* **117** S39–S58. [MR3219149](#)
- DECKER, S. H., SPOHN, C., ORTIZ, N. R. and HEDBERG, E. (2014). Criminal stigma, race, gender, and employment: An expanded assessment of the consequences of imprisonment for employment. (Final report to the National Institute of Justice, 2010-MU-MU-0004.) National Institute of Justice, Washington, DC.

- ERTEFAIE, A., SMALL, D. S. and ROSENBAUM, P. R. (2018). Quantitative evaluation of the trade-off of strengthened instruments and sample size in observational studies. *J. Amer. Statist. Assoc.* **113** 1122–1134.
- ERTEFAIE, A., SMALL, D., FLORY, J. and HENNESSY, S. (2016a). Selection bias when using instrumental variable methods to compare two treatments but more than two treatments are available. *Int. J. Biostat.* **12** 219–232. [MR3505695](#)
- ERTEFAIE, A., SMALL, D. S., FLORY, J. and HENNESSY, S. (2016b). A sensitivity analysis to assess bias due to selecting subjects based on treatment received. *Epidemiology* **27** e5–e7.
- ERTEFAIE, A., SMALL, D. S., FLORY, J. H. and HENNESSY, S. (2017). A tutorial on the use of instrumental variables in pharmacoepidemiology. *Pharmacoepidemiol. Drug Saf.* **26** 357–367.
- ERTEFAIE, A., NGUYEN, A., HARDING, D., MORENOFF, J. and YANG, W. (2018). Supplement to “Instrumental variable analysis with censored data in the presence of many weak instruments: Application to the effect of being sentenced to prison on time to employment.” DOI:[10.1214/18-AOAS1174SUPP](#).
- FULLER, W. A. (1977). Some properties of a modification of the limited information estimator. *Econometrica* **45** 939–953. [MR0451608](#)
- HAANELMO, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12. [MR0007954](#)
- HAHN, J., HAUSMAN, J. and KUERSTEINER, G. (2004). Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations. *Econom. J.* **7** 272–306. [MR2076636](#)
- HAHN, J. and INOUE, A. (2002). A Monte Carlo comparison of various asymptotic approximations to the distribution of instrumental variables estimators. *Econometric Rev.* **21** 309–336. [MR1944978](#)
- HANSEN, C., HAUSMAN, J. and NEWEY, W. (2008). Estimation with many instrumental variables. *J. Bus. Econom. Statist.* **26** 398–422. [MR2459342](#)
- HAUSMAN, C. H. J. and NEWEY, W. K. (2004). Many instruments, weak instruments, and microeconomic practice.
- HERNÁN, M. A. and ROBINS, J. M. (2006). Instruments for causal inference: An epidemiologist’s dream? *Epidemiology* **17** 360–372.
- HERNÁN, M. A., COLE, S. R., MARGOLICK, J., COHEN, M. and ROBINS, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiol. Drug Saf.* **14** 477–491.
- IMBENS, G. W. and ROSENBAUM, P. R. (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *J. Roy. Statist. Soc. Ser. A* **168** 109–126. [MR2113230](#)
- INOUE, A. and SOLON, G. (2010). Two-sample instrumental variables estimators. *Rev. Econ. Stat.* **92** 557–561.
- JOFFE, M. M. (2001). Administrative and artificial censoring in censored regression models. *Stat. Med.* **20** 2287–2304.
- KADANE, J. B. and ANDERSON, T. W. (1977). A comment on the test of overidentifying restrictions. *Econometrica* **45** 1027–1031. [MR0483260](#)
- KAFFO, M. and WANG, W. (2017). On bootstrap validity for specification testing with many weak instruments. *Econom. Lett.* **157** 107–111. [MR3669092](#)
- KANG, H., CAI, T. T. and SMALL, D. S. (2015). Robust confidence intervals for causal effects with possibly invalid instruments. ArXiv Preprint [arXiv:1504.03718](#).
- KILGORE, J. (2015). *Understanding Mass Incarceration: A People’s Guide to the Key Civil Rights Struggle of Our Time*. The New Press, New York.
- KLEIBERGEN, F. (2007). Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices and identification statistics. *J. Econometrics* **139** 181–216. [MR2380764](#)
- KLING, J. R. (2006). Incarceration length, employment, and earnings. Technical report, National Bureau of Economic Research, New York.

- KOLESÁR, M., CHETTY, R., FRIEDMAN, J., GLAESER, E. and IMBENS, G. W. (2015). Identification and inference with many invalid instruments. *J. Bus. Econom. Statist.* **33** 474–484. [MR3416595](#)
- LI, J., FINE, J. and BROOKHART, A. (2015). Instrumental variable additive hazards models. *Biometrics* **71** 122–130. [MR3335356](#)
- LI, Y., LEE, Y., WOLFE, R. A., MORGENSTERN, H., ZHANG, J., PORT, F. K. and ROBINSON, B. M. (2015). On a preference-based instrumental variable approach in reducing unmeasured confounding-by-indication. *Stat. Med.* **34** 1150–1168. [MR3322741](#)
- LOEFFLER, C. E. (2013). Does imprisonment alter the life course? Evidence on crime and employment from a natural experiment. *Criminology* **51** 137–166.
- MACKENZIE, T. A., TOSTESON, T. D., MORDEN, N. E., STUKEL, T. A. and O'MALLEY, A. J. (2014). Using instrumental variables to estimate a Cox's proportional hazards regression subject to additive confounding. *Health Serv. Outcomes Res. Methodol.* **14** 54–68.
- MAGDALINOS, M. A. (1990). The classical principles of testing using instrumental variables estimates. *J. Econometrics* **44** 241–279. [MR1060648](#)
- MAGDALINOS, M. A. (1994). Testing instrument admissibility: Some refined asymptotic results. *Econometrica* **62** 373–403. [MR1281696](#)
- MARK, S. D. and ROBINS, J. M. (1993). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Stat. Med.* **12** 1605–1628.
- MAUER, M. (2001). The causes and consequences of prison growth in the United States. *Punishment & Society* **3** 9–20.
- NAGIN, D. S., CULLEN, F. T. and JONSON, C. L. (2009). Imprisonment and reoffending. *Crime Justice* **38** 115–200.
- NEWWEY, W. K. and WINDMEIJER, F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica* **77** 687–719. [MR2531359](#)
- PAGER, D. (2003). The mark of a criminal record 1. *Amer. J. Sociol.* **108** 937–975.
- PAGER, D. (2008). *Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration*. Univ. Chicago Press, Chicago, IL.
- ROBINS, J. M. (1993). Analytic methods for estimating HIV treatment and cofactor effects. In *Methodological Issues of AIDS Mental Health Research* (R. Kessler, ed.). Plenum Publishing, New York.
- ROBINS, J. M. (1997). Structural nested failure time models. In *The Encyclopedia of Biostatistics* (T. Colton, ed.). In the section *Survival Analysis* (P. K. Andersen and N. Keiding, Section eds.) 4372–4389. Wiley, Chichester.
- ROBINS, J. M. and TSIATIS, A. A. (1991). Correcting for noncompliance in randomized trials using rank preserving structural failure time models. *Comm. Statist. Theory Methods* **20** 2609–2631. [MR1144866](#)
- RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. [MR0472152](#)
- SMALL, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J. Amer. Statist. Assoc.* **102** 1049–1058. [MR2411664](#)
- SMALL, D. S. and ROSENBAUM, P. R. (2008). War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *J. Amer. Statist. Assoc.* **103** 924–933. [MR2528819](#)
- SOLOMON, B. J., MOK, T., KIM, D.-W., WU, Y.-L., NAKAGAWA, K., MEKHAIL, T., FELIP, E., CAPPUZZO, F., PAOLINI, J., USARI, T. et al. (2014). First-line crizotinib versus chemotherapy in ALK-positive lung cancer. *N. Engl. J. Med.* **371** 2167–2177.
- SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. [MR1092986](#)
- STAIGER, D. O. and STOCK, J. H. (1994). *Instrumental variables regression with weak instruments*. National Bureau of Economic Research, Cambridge, MA.

- STAIGER, D. and STOCK, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65** 557–586. [MR1445622](#)
- STEFFENSMEIER, D., ULMER, J. and KRAMER, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology* **36** 763–798.
- STOCK, J. H., WRIGHT, J. H. and YOGO, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econom. Statist.* **20** 518–529. [MR1973801](#)
- STOCK, J. H. and YOGO, M. (2005). Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models* 80–108. Cambridge Univ. Press, Cambridge. [MR2232140](#)
- SWANSON, S. A., ROBINS, J. M., MILLER, M. and HERNÁN, M. A. (2015). Selecting on treatment: A pervasive form of bias in instrumental variable analyses. *Am. J. Epidemiol.* **181** 191–197.
- TCHETGEN, E. J. T., WALTER, S., VANSTEELANDT, S., MARTINUSSEN, T. and GLYMOUR, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology* **26** 402.
- TERZA, J. V., BASU, A. and RATHOUZ, P. J. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *J. Health Econ.* **27** 531–543.
- TRAVIS, J., WESTERN, B. and REDBURN, F. S. (2014). *The Growth of Incarceration in the United States: Exploring Causes and Consequences*. National Academies Press, Washington, DC.
- TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York. [MR2233926](#)
- TURNER, K. and WILDEMAN, C. (2015). Detrimental for some? Heterogeneous effects of maternal incarceration on child wellbeing. *Criminol. Public Policy* **14** 125–156.
- WANG, X., JIANG, Y., ZHANG, N. and SMALL, D. (2018). Sensitivity analysis and power for instrumental variable studies. *Biometrics*. DOI: [10.1111/biom.12873](#).
- WEIMAN, D. F., STOLL, M. A. and BUSHWAY, S. (2007). The regime of mass-incarceration: A labor-market perspective. In *Barriers to Reentry?: The Labor Market for Released Prisoners in Post-Industrial America* (S. Bushway, M. A. Stoll and D. F. Weiman, eds.) 29–79. Russell Sage Foundation, New York.
- WESTERN, B. (2002). The impact of incarceration on wage mobility and inequality. *Am. Sociol. Rev.* 526–546.
- WRIGHT, S. (1934). The method of path coefficients. *Ann. Math. Stat.* **5** 161–215.

VARIATIONAL INFERENCE FOR PROBABILISTIC POISSON PCA¹

BY JULIEN CHIQUET, MAHENDRA MARIADASSOU AND STÉPHANE ROBIN

AgroParisTech, INRA and Université Paris-Saclay

Many application domains, such as ecology or genomics, have to deal with multivariate non-Gaussian observations. A typical example is the joint observation of the respective abundances of a set of species in a series of sites aiming to understand the covariations between these species. The Gaussian setting provides a canonical way to model such dependencies but does not apply in general. We consider here the multivariate exponential family framework for which we introduce a generic model with multivariate Gaussian latent variables. We show that approximate maximum likelihood inference can be achieved via a variational algorithm for which gradient descent easily applies. We show that this setting enables us to account for covariates and offsets. We then focus on the case of the Poisson-lognormal model in the context of community ecology. We demonstrate the efficiency of our algorithm on microbial ecology datasets. We illustrate the importance of accounting for the effects of covariates to better understand interactions between species.

REFERENCES

- ACHARYA, A., GHOSH, J. and ZHOU, M. (2015). Nonparametric Bayesian factor analysis for dynamic count matrices. In *AISTATS*.
- AITCHISON, J. and HO, C.-H. (1989). The multivariate Poisson-log normal distribution. *Biometrika* **76** 643–653. [MR1041409](#)
- ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley-Interscience, Hoboken, NJ. [MR1990662](#)
- 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** 719–25.
- CAO, Y. and XIE, Y. (2015). Poisson matrix completion. In *2015 IEEE International Symposium on Information Theory (ISIT)* 1841–1845. IEEE, New York.
- CHEN, J., KING, E., DEEK, R., WEI, Z., YU, Y., GRILL, D. and BALLMAN, K. (2018). An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics* **34** 643–651.
- COLLINS, M., DASGUPTA, S. and SCHAPIRE, R. E. (2001). A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems* 617–624.
- 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](#)
- ECKART, C. and YOUNG, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1** 211–218.
- GLOOR, G. B., MACKLAIM, J. M., PAWLOWSKY-GLAHN, V. and EGOZCUE, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8** 2224.

Key words and phrases. Probabilistic PCA, Poisson-lognormal model, count data, variational inference.

- HALL, P., ORMEROD, J. T. and WAND, M. P. (2011). Theory of Gaussian variational approximation for a Poisson mixed model. *Statist. Sinica* **21** 369–389. [MR2796867](#)
- IZSÁK, R. (2008). Maximum likelihood fitting of the Poisson lognormal distribution. *Environ. Ecol. Stat.* **15** 143–156. [MR2399077](#)
- JAAKKOLA, T. S. and JORDAN, M. I. (2000). Bayesian parameter estimation via variational methods. *Stat. Comput.* **10** 25–37.
- JAKUSCHKIN, B., FIEVET, V., SCHWALLER, L., FORT, T., ROBIN, C. and VACHER, C. (2016). Deciphering the pathobiome: Intra-and interkingdom interactions involving the pathogen *Erysiphe alphitoides*. *Microb. Ecol.* **72** 870–880.
- JOHNSON, S. G. (2011). The NLOpt nonlinear-optimization package. Available at <http://ab-initio.mit.edu/nlopt>.
- JOHNSON, N. L., KOTZ, S. and BALAKRISHNAN, N. (1997). *Discrete Multivariate Distributions*. Wiley, New York. [MR1429617](#)
- KARLIS, D. (2005). EM algorithm for mixed Poisson and other discrete distributions. *Astin Bull.* **35** 3–24. [MR2142681](#)
- LAFOND, J. (2015). Low rank matrix completion with exponential family noise. Preprint. Available at [arXiv:1502.06919](https://arxiv.org/abs/1502.06919).
- LANDGRAF, A. J. (2015). Generalized principal component analysis: Dimensionality reduction through the projection of natural parameters. Ph.D. thesis, Ohio State Univ., Columbus, OH. [MR3450322](#)
- LEE, D. D. and SEUNG, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems* 556–562.
- LI, J. and TAO, D. (2010). Simple exponential family PCA. In *AISTATS* 453–460.
- LITTLE, R. J. A. and RUBIN, D. B. (2014). *Statistical Analysis with Missing Data*, Wiley-Interscience, Hoboken, NJ. [MR1925014](#)
- LIU, L. T., DOBRIBAN, E. and SINGER, A. (2016). ePCA: High dimensional exponential family PCA. Preprint. Available at [arXiv:1611.05550](https://arxiv.org/abs/1611.05550).
- MACH, N., BERRI, M., ESTELLÉ, J., LEVENEZ, F., LEMONNIER, G., DENIS, C., LEPLAT, J.-J., CHEVALEYRE, C., BILLON, Y., DORÉ, J., ROGEL-GAILLARD, C. and LEPAGE, P. (2015). Early-life establishment of the swine gut microbiome and impact on host phenotypes. *Environ. Microbiol. Rep.* **7** 554–569.
- MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, New York. [MR0560319](#)
- MINKA, T. P. (2000). Automatic choice of dimensionality for PCA. In *NIPS* **13** 598–604.
- MOHAMED, S., GHAHRAMANI, Z. and HELLER, K. A. (2009). Bayesian exponential family PCA. In *Advances in Neural Information Processing Systems* 1089–1096.
- NELSON, J. F. (1985). Multivariate gamma-Poisson models. *J. Amer. Statist. Assoc.* **80** 828–834.
- PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. and FLANNERY, B. P. (1989). *Numerical Recipes: The Art of Scientific Computing*. Code CD-ROM v 2.06 with UNIX Single-Screen License, 3rd ed. Cambridge Univ. Press, Cambridge. [MR1414682](#)
- R DEVELOPMENT CORE TEAM (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at <http://www.R-project.org>.
- ROYLE, J. A. and WIKLE, C. K. (2005). Efficient statistical mapping of avian count data. *Environ. Ecol. Stat.* **12** 225–243. [MR2144403](#)
- SALMON, J., HARMANY, Z., DELEDALLE, C.-A. and WILLETT, R. (2014). Poisson noise reduction with non-local PCA. *J. Math. Imaging Vision* **48** 279–294. [MR3152105](#)
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. [MR0468014](#)
- SMETS, W., LEFF, J. W., BRADFORD, M. A., MCCULLEY, R. L., LEBEER, S. and FIERER, N. (2015). A method for simultaneous measurement of soil bacterial abundances and community composition via 16s rRNA gene sequencing. *PeerJ PrePrints* **3** e1318v1.

- SRIVASTAVA, S. and CHEN, L. (2010). A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res.* **38** e170–e170.
- SVANBERG, K. (2002). A class of globally convergent optimization methods based on conservative convex separable approximations. *SIAM J. Optim.* **12** 555–573. [MR1885575](#)
- TIPPING, M. E. and BISHOP, C. M. (1999). Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 611–622. [MR1707864](#)
- TSILIMIGRAS, M. C. B. and FODOR, A. A. (2016). Compositional data analysis of the microbiome: Fundamentals, tools, and challenges. *Ann. Epidemiol.* **26** 330–335.
- VANDEPUTTE, D., KATHAGEN, G., D’HOE, K., VIEIRA-SILVA, S., VALLES-COLOMER, M., SABINO, J., WANG, J., TITO, R. Y., DE COMMER, L., DARZI, Y. et al. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551** 507–511.
- WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.
- WICKHAM, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. Springer, New York. Available at <http://ggplot2.org>.
- WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.
- YPMA, J. (2017). R interface to NLOpt, v. 1.0.4. Available at <https://github.com/jyypma/nloptr>.
- ZHOU, M. (2016). Nonparametric Bayesian negative binomial factor analysis. Preprint. Available at [arXiv:1604.07464](https://arxiv.org/abs/1604.07464).
- ZHOU, M., HANNAH, L., DUNSON, D. B. and CARIN, L. (2012). Beta-negative binomial process and Poisson factor analysis. In *AISTATS* **22** 1462–1471.

SPLIT-DOOR CRITERION: IDENTIFICATION OF CAUSAL EFFECTS THROUGH AUXILIARY OUTCOMES

BY AMIT SHARMA, JAKE M. HOFMAN AND DUNCAN J. WATTS

Microsoft Research

We present a method for estimating causal effects in time series data when fine-grained information about the outcome of interest is available. Specifically, we examine what we call the *split-door* setting, where the outcome variable can be split into two parts: one that is potentially affected by the cause being studied and another that is independent of it, with both parts sharing the same (unobserved) confounders. We show that under these conditions, the problem of identification reduces to that of testing for independence among observed variables, and propose a method that uses this approach to automatically find subsets of the data that are causally identified. We demonstrate the method by estimating the causal impact of Amazon’s recommender system on traffic to product pages, finding thousands of examples within the dataset that satisfy the split-door criterion. Unlike past studies based on natural experiments that were limited to a single product category, our method applies to a large and representative sample of products viewed on the site. In line with previous work, we find that the widely-used click-through rate (CTR) metric overestimates the causal impact of recommender systems; depending on the product category, we estimate that 50–80% of the traffic attributed to recommender systems would have happened even without any recommendations. We conclude with guidelines for using the split-door criterion as well as a discussion of other contexts where the method can be applied.

REFERENCES

- AGRESTI, A. (1992). A survey of exact inference for contingency tables. *Statist. Sci.* **7** 131–177. [MR1173420](#)
- AGRESTI, A. (2001). Exact inference for categorical data: Recent advances and continuing controversies. *Stat. Med.* **20** 2709–2722.
- ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.
- CARMİ, E., OESTREICHER-SINGER, G. and SUNDARARAJAN, A. (2012). Is Oprah contagious? Identifying demand spillovers in online networks. NET Institute Working Paper 10-18.
- CARNEGIE, N. B., HARADA, M. and HILL, J. L. (2016). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness* **9** 395–420.
- CATTANEO, M. D., FRANDSEN, B. R. and TITIUNIK, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference* **3** 1–24.

Key words and phrases. Causal inference, data mining, causal graphical model, natural experiment, recommendation systems.

- CATTANEO, M. D., TITIUNIK, R. and VAZQUEZ-BARE, G. (2017). Comparing inference approaches for RD designs: A reexamination of the effect of head start on child mortality. *Journal of Policy Analysis and Management* **36** 643–681.
- DE SIQUEIRA SANTOS, S., TAKAHASHI, D. Y., NAKATA, A. and FUJITA, A. (2014). A comparative study of statistical methods used to identify dependencies between gene expression signals. *Brief. Bioinform.* **15** 906–918.
- DELONGCHAMP, R. R., BOWYER, J. F., CHEN, J. J. and KODELL, R. L. (2004). Multiple-testing strategy for analyzing cDNA array data on gene expression. *Biometrics* **60** 774–782. [MR2089454](#)
- DUNNING, T. (2012). *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge Univ. Press, Cambridge.
- FARCOMENI, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat. Methods Med. Res.* **17** 347–388. [MR2526659](#)
- FISKE, S. T. and HAUSER, R. M. (2014). Protecting human research participants in the age of big data. *Proc. Natl. Acad. Sci. USA* **111** 13675–13676.
- FLAXMAN, S., GOEL, S. and RAO, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opin. Q.* **80** 298–320.
- GRAU, J. (2009). Personalized product recommendations: Predicting shoppers' needs. *EMarketer*.
- GROSSE-WENTRUP, M., JANZING, D., SIEGEL, M. and SCHÖLKOPF, B. (2016). Identification of causal relations in neuroimaging data with latent confounders: An instrumental variable approach. *NeuroImage* **125** 825–833.
- HARDING, D. J. (2009). Collateral consequences of violence in disadvantaged neighborhoods. *Soc. Forces* **88** 757–784.
- IMBENS, G. W. (2010). Better LATE than nothing. *J. Econ. Lit.* **48**.
- IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—For Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. [MR3309951](#)
- JENSEN, D. D., FAST, A. S., TAYLOR, B. J. and MAIER, M. E. (2008). Automatic identification of quasi-experimental designs for discovering causal knowledge. In *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining* 372–380.
- KANG, H., ZHANG, A., CAI, T. T. and SMALL, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *J. Amer. Statist. Assoc.* **111** 132–144. [MR3494648](#)
- LEWIS, R. A., RAO, J. M. and REILEY, D. H. (2011). Here, there, and everywhere: Correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the 20th International Conference on World Wide Web* 157–166. ACM, New York.
- LIANG, K. and NETTLETON, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 163–182. [MR2885844](#)
- LYDERSEN, S., PRADHAN, V., SENCHAUDHURI, P. and LAAKE, P. (2007). Choice of test for association in small sample unordered $r \times c$ tables. *Stat. Med.* **26** 4328–4343. [MR2405356](#)
- MEALLI, F. and PACINI, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *J. Amer. Statist. Assoc.* **108** 1120–1131. [MR3174688](#)
- MORGAN, S. L. and WINSHIP, C. (2014). *Counterfactuals and Causal Inference*. Cambridge Univ. Press, Cambridge.
- MULPURU, S. (2006). What you need to know about third-party recommendation engines. *Forrester Research*.
- NETTLETON, D., HWANG, J. T. G., CALDO, R. A. and WISE, R. P. (2006). Estimating the number of true null hypotheses from a histogram of p values. *J. Agric. Biol. Environ. Stat.* **11** 337.
- PANINSKI, L. (2003). Estimation of entropy and mutual information. *Neural Comput.* **15** 1191–1253.
- PEARL, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. [MR2548166](#)
- PETHEL, S. D. and HAHS, D. W. (2014). Exact test of independence using mutual information. *Entropy* **16** 2839–2849.

- PHAN, T. Q. and AIROLDI, E. M. (2015). A natural experiment of social network formation and dynamics. *Proc. Natl. Acad. Sci. USA* **112** 6595–6600.
- RICCI, F., ROKACH, L. and SHAPIRA, B. (2011). *Introduction to Recommender Systems Handbook*. Springer, Berlin.
- ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York. [MR2561612](#)
- ROSENZWEIG, M. R. and WOLPIN, K. I. (2000). Natural “natural experiments” in economics. *J. Econ. Lit.* **38** 827–874.
- RUBIN, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge Univ. Press, Cambridge. [MR2307965](#)
- SHARMA, A., HOFMAN, J. M. and WATTS, D. J. (2015). Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the 16th ACM Conference on Economics and Computation* 453–470.
- SHARMA, A., HOFMAN, J. M. and WATTS, D. J. (2018). Supplement to “Split-door criterion: Identification of causal effects through auxiliary outcomes.” DOI:[10.1214/18-AOAS1179SUPP](#).
- SPIRITES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. MIT Press, Cambridge, MA. [MR1815675](#)
- STEUER, R., KURTHS, J., DAUB, C. O., WEISE, J. and SELBIG, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* **18** S231–S240.
- STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. [MR1924302](#)
- STOREY, J. D. and TIBSHIRANI, R. (2003). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data*. 272–290. Springer, New York. [MR2001400](#)
- STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. [MR2741812](#)
- SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. [MR2382665](#)
- VANDERWEELE, T. J. and ARAH, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* **22** 42–52.

The Annals of Applied Statistics

Next Issues

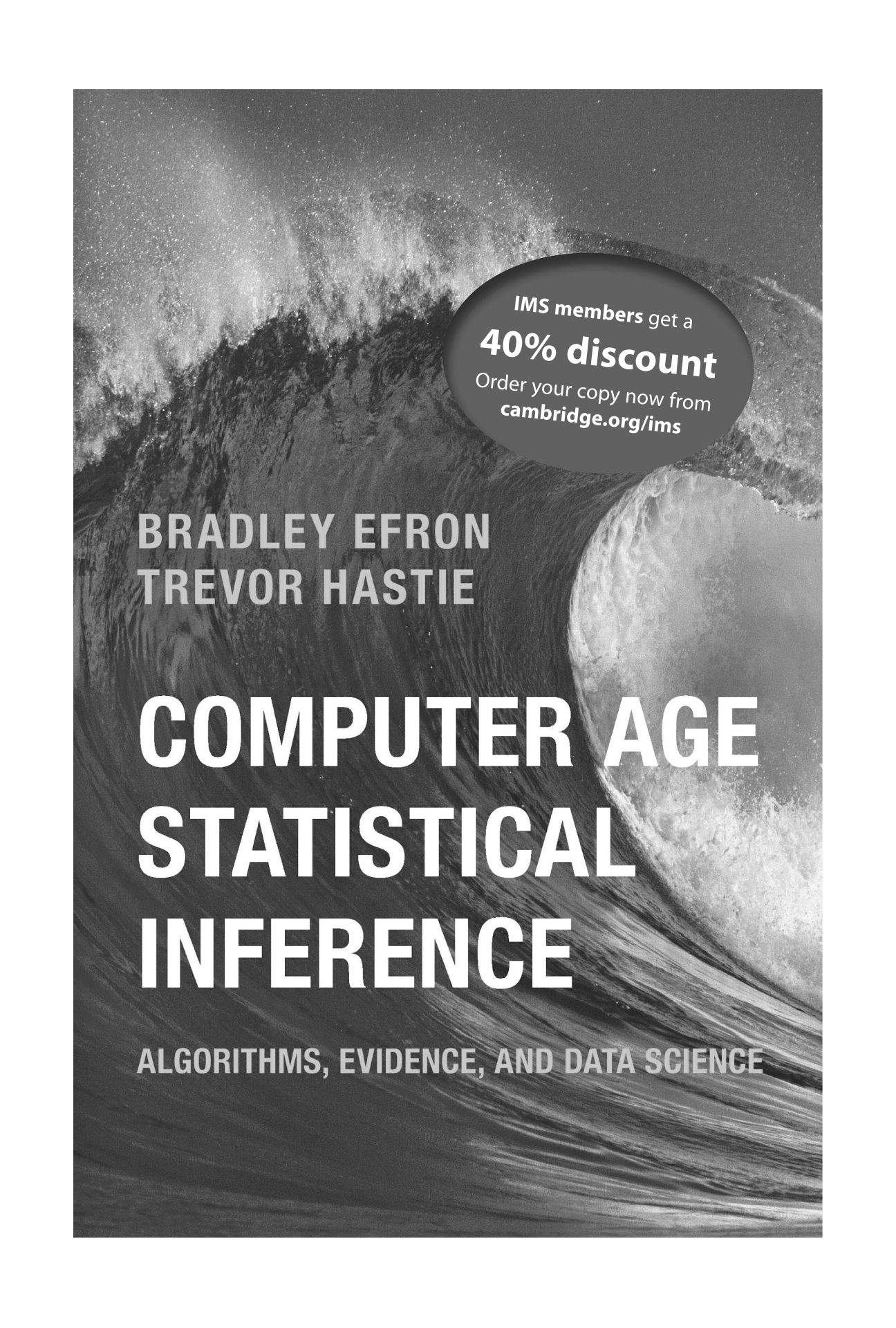
- Epigenome-wide analyses of sparse mediation effects under composite null hypotheses YEN-TSUNG HUANG
- Joint Bayesian semiparametric regression analysis of recurrent adverse events and survival in esophageal cancer patients JUHEE LEE, PETER F. THALL AND STEVEN H. LIN
- Ground-level ozone: Evidence of increasing serial dependence in the extremes
DEBBIE J. DUPUIS AND LUCA TRAPIN
- Multilayer knockoff filter: Controlled variable selection at multiple resolutions
EUGENE KATSEVICH AND CHIARA SABATTI
- A penalized regression model for the joint estimation of eQTL associations and gene network structure
MICOL MARCHETTI-BOWICK, YAOLIANG YU, WEI WU AND ERIC POE XING
- Joint mean and covariance modeling of multiple health outcome measures
XIAOYUE NIU AND PETER D. HOFF
- Bayesian latent hierarchical model for transcriptomic meta-analysis to detect biomarkers with clustered meta-patterns of differential expression signals
ZHIGUANG HUO, CHI SONG AND GEORGE TSENG
- Modeling within-household associations in household panel studies
FIONA STEELE, PAUL CLARKE AND JOUNI KUHA
- A Bayesian race model for response times under cyclic stimulus discriminability
DEBORAH KUNKEL, KEVIN POTTER, PETER F. CRAIGMILE,
MARIO PERUGGIA AND TRISHA VAN ZANDT
- Common and individual structure of brain networks
LU WANG, ZHENGWU ZHANG AND DAVID DUNSON
- Fréchet estimation of time-varying covariance matrices from sparse data, with application to the regional co-evolution of myelination in the developing brain
ALEXANDER PETERSEN, SEAN DEONI AND HANS-GEORG MÜLLER
- The role of mastery learning in intelligent tutoring systems: Principal stratification on a latent variable ADAM C. SALES AND JOHN F. PANE
- Clonality: Point estimation
LU TIAN, YI LIU, SCOTT BOYD, ANDREW FIRE AND RICHARD OLSHEN
- Capturing heterogeneity of covariate effects in hidden subpopulations in the presence of censoring and large number of covariates
FARHAD SHOKOOHI, ABBAS KHALILI, MASOUD ASGHARIAN AND SHILI LIN
- Bayesian analysis of infant's growth dynamics with in utero exposure to environmental toxicants JONGGYU BAEK, BIN ZHU AND PETER X. K. SONG
- Dynamics of homelessness in urban America CHRIS GLYNN AND EMILY B. FOX
- An algorithm for removing sensitive information: Application to race-independent recidivism prediction JAMES JOHNDROW AND KRISTIAN LUM
- Development of a common patient assessment scale across the continuum of care: A nested multiple imputation approach CHENYANG GU AND ROEE GUTMAN
- A Bayesian Mallows approach to non-transitive pair comparison data: How human are sounds?
MARTA CRISPINO, ELJA ARJAS, VALERIA VITELLI,
NATASHA BARRETT AND ARNOLDO FRIGESSI

Continued

The Annals of Applied Statistics

Next Issues–Continued

- Nonstationary spatial prediction of soil organic carbon: Implications for stock assessment
decision making MARK D. RISSE, CATHERINE A. CALDER, VERONICA J.
BERROCAL AND CANDACE BERRETT
- Prediction models for network-linked data TIANXI LI, ELIZAVETA LEVINA AND JI ZHU
- Causal inference in the context of an error prone exposure: Air pollution and mortality
XIAO WU, DANIELLE BRAUN, MARIANTHI-ANNA KIOUMOURTZOGLU,
CHRISTINE CHOIRAT, QIAN DI AND FRANCESCA DOMINICI
- Modeling biomarker ratios with gamma distributed components
MORITZ BERGER, MICHAEL WAGNER AND MATTHIAS SCHMID
- Bayesian hidden Markov tree models for clustering genes with shared evolutionary history
YANG LI, SHAOYANG NING, SARAH E. CALVO, VAMSI K. MOOTHA AND JUN S. LIU
- Sequential Dirichlet process mixtures of multivariate skew t-distributions for model-based
clustering of flow cytometry data BORIS P. HEJBLUM, CHARIFF ALKHASSIM,
RAPHAEL GOTTARDO, FRANÇOIS CARON AND RODOLPHE THIÉBAUT
- Compositional mediation analysis for microbiome studies
MICHAEL SOHN AND HONGZHE LI
- Learning algorithms to evaluate forensic glass evidence
SOYOUNG PARK AND ALICIA CARRIQUIRY
- A hierarchical multivariate spatio-temporal model for clustered climate data with annual cycles
GIANLUCA MASTRANTONIO, GIOVANNA JONA LASINIO, ALESSIO POLLICE, GIULIA
CAPOTORTI, LORENZO TEODONIO, GIULIO GENOVA AND CARLO BLASI
- Graphical models for zero-inflated single cell gene expression
ANDREW MCDAVID, RAPHAEL GOTTARDO, NOAH SIMON AND MATHIAS DRTON
- Modelling multilevel spatial behaviour in binary-mark muscle fibre configurations
TILMAN DAVIES, MATTHEW SCHOFIELD, JON CORNWALL AND PHILIP SHEARD
- Extended sensitivity analysis for heterogeneous unmeasured confounding with an application to
sibling studies of returns to education . . COLIN B. FOGARTY AND RAIDEN B. HASEGAWA
- Fused comparative intervention scoring for heterogeneity of longitudinal intervention effects
JARED DAVIS HULING, MENGANG YU, AND MAUREEN A. SMITH
- Modelling ocean temperatures from bio-probes under preferential sampling
DANIEL DINSDALE AND MATÍAS SALIBIÁN-BARRERA
- Climate inference on daily rainfall across the Australian continent, 1876-2015
MICHAEL BERTOLACCI, EDWARD JERROLD CRIPPS, ORI ROSEN,
JOHN LAU AND SALLY ANN WOOD
- Complete spatial model calibration YEN-NING HUANG, BRIAN J. REICH,
MONTSERRAT FUENTES AND A. SANKARASUBRAMANIAN



IMS members get a
40% discount

Order your copy now from
cambridge.org/ims

BRADLEY EFRON
TREVOR HASTIE

COMPUTER AGE STATISTICAL INFERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE