

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

- Estimating the rate constant from biosensor data via an adaptive variational Bayesian approach YE ZHANG, ZHIGANG YAO, PATRIK FORSSÉN AND TORGNÝ FORNSTEDT 2011
- Estimating abundance from multiple sampling capture-recapture data via a multi-state multi-period stopover model HANNAH WORTHINGTON, RACHEL MCCREA, RUTH KING AND RICHARD GRIFFITHS 2043
- Robust elastic net estimators for variable selection and identification of proteomic biomarkers GABRIELA V. COHEN FREUE, DAVID KEPPLINGER, MATÍAS SALIBIÁN-BARRERA AND EZEQUIEL SMUCLER 2065
- Statistical inference for partially observed branching processes with application to cell lineage tracking of *in vivo* hematopoiesis
JASON XU, SAMSON KOELLE, PETER GUTTORP, CHUANFENG WU, CYNTHIA DUNBAR, JANIS L. ABKOWITZ AND VLADIMIR N. MININ 2091
- Fire seasonality identification with multimodality tests
JOSE AMEIJERAS-ALONSO, AKLI BENALI, ROSA M. CRUJEIRAS, ALBERTO RODRÍGUEZ-CASAL AND JOSÉ M. C. PEREIRA 2120
- Joint model of accelerated failure time and mechanistic nonlinear model for censored covariates, with application in HIV/AIDS HONGBIN ZHANG AND LANG WU 2140
- Prediction of small area quantiles for the conservation effects assessment project using a mixed effects quantile regression model EMILY BERG AND DANHYANG LEE 2158
- Microsimulation model calibration using incremental mixture approximate Bayesian computation CAROLYN M. RUTTER, JONATHAN OZIK, MARIA DE YOREO AND NICHOLSON COLLIER 2189
- Principal nested shape space analysis of molecular dynamics data IAN L. DRYDEN, KWANG-RAE KIM, CHARLES A. LAUGHTON AND HUILING LE 2213
- Spatial modeling of trends in crime over time in Philadelphia
CECILIA BALOCCHI AND SHANE T. JENSEN 2235
- Fitting a deeply nested hierarchical model to a large book review dataset using a moment-based estimator NINGSHAN ZHANG, KYLE SCHMAUS AND PATRICK O. PERRY 2260
- Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways FEDERICO CASTELLETTI AND GUIDO CONSONNI 2289
- A latent discrete Markov random field approach to identifying and classifying historical forest communities based on spatial multivariate tree species counts
STEPHEN BERG, JUN ZHU, MURRAY K. CLAYTON, MONIKA E. SHEA AND DAVID J. MLADENOFF 2312
- A nonparametric spatial test to identify factors that shape a microbiome
SUSHEELA P. SINGH, ANA-MARIA STAICU, ROBERT R. DUNN, NOAH FIERER AND BRIAN J. REICH 2341

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—Continued from front cover

- Predicting paleoclimate from compositional data using multivariate Gaussian process
inverse prediction JOHN R. TIPTON, MEVIN B. HOOTEN, CONNOR NOLAN,
ROBERT K. BOOTH AND JASON MCLACHLAN 2363
- Propensity score weighting for causal inference with multiple treatments
FAN LI AND FAN LI 2389
- Outline analyses of the called strike zone in Major League Baseball
DALE L. ZIMMERMAN, JUN TANG AND RUI HUANG 2416
- Empirical Bayes analysis of RNA sequencing experiments with auxiliary information
KUN LIANG 2452
- New formulation of the logistic-Gaussian process to analyze trajectory tracking
data GIANLUCA MASTRANTONIO, CLARA GRAZIAN, SARA MANCINELLI
AND ENRICO BIBBONA 2483
- A simple, consistent estimator of SNP heritability from genome-wide association
studies ARMIN SCHWARTZMAN, ANDREW J. SCHORK, RONG ZABLOCKI
AND WESLEY K. THOMPSON 2509
- A hierarchical curve-based approach to the analysis of manifold data
LIBERTY VITTERT, ADRIAN W. BOWMAN AND STANISLAV KATINA 2539
- Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods:
A winning solution to the NIJ “Real-Time Crime Forecasting Challenge”
SETH FLAXMAN, MICHAEL CHIRICO, PAU PEREIRA AND CHARLES LOEFFLER 2564
- On Bayesian new edge prediction and anomaly detection in computer networks
SILVIA METELLI AND NICHOLAS HEARD 2586
- Bayesian indicator variable selection to incorporate hierarchical overlapping group
structure in multi-omics applications . . . LI ZHU, ZHIGUANG HUO, TIANZHOU MA,
STEFFI OESTERREICH AND GEORGE C. TSENG 2611
- Hierarchical infinite factor models for improving the prediction of surgical complications
for geriatric patients ELIZABETH LORENZI, RICARDO HENAO
AND KATHERINE HELLER 2637
- A general theory for preferential sampling in environmental networks
JOE WATSON, JAMES V. ZIDEK AND GAVIN SHADDICK 2662

THE ANNALS OF APPLIED STATISTICS

Vol. 13, No. 4, pp. 2011–2700 December 2019

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

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

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

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

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

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

IMS PUBLICATIONS

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

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

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

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

Statistical Science. *Editor:* 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 13, Number 4, December 2019. 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.

ESTIMATING THE RATE CONSTANT FROM BIOSENSOR DATA VIA AN ADAPTIVE VARIATIONAL BAYESIAN APPROACH

BY YE ZHANG^{*,1}, ZHIGANG YAO^{†,2,4}, PATRIK FORSSÉN^{‡,3} AND
 TORGNY FORNSTEDT^{‡,3}

Chemnitz University of Technology^{*}, *National University of Singapore*[†] and
Karlstad University[‡]

The means to obtain the rate constants of a chemical reaction is a fundamental open problem in both science and the industry. Traditional techniques for finding rate constants require either chemical modifications of the reactants or indirect measurements. The rate constant map method is a modern technique to study binding equilibrium and kinetics in chemical reactions. Finding a rate constant map from biosensor data is an ill-posed inverse problem that is usually solved by regularization. In this work, rather than finding a deterministic regularized rate constant map that does not provide uncertainty quantification of the solution, we develop an adaptive variational Bayesian approach to estimate the distribution of the rate constant map, from which some intrinsic properties of a chemical reaction can be explored, including information about rate constants. Our new approach is more realistic than the existing approaches used for biosensors and allows us to estimate the dynamics of the interactions, which are usually hidden in a deterministic approximate solution. We verify the performance of the new proposed method by numerical simulations, and compare it with the Markov chain Monte Carlo algorithm. The results illustrate that the variational method can reliably capture the posterior distribution in a computationally efficient way. Finally, the developed method is also tested on the real biosensor data (parathyroid hormone), where we provide two novel analysis tools—the thresholding contour map and the high order moment map—to estimate the number of interactions as well as their rate constants.

REFERENCES

- ALTSCHUH, D., BJÖKELUND, H., STRANDGÅRD, J., CHOULIER, L., MALMQVIST, M. and ANDERSSON, K. (2012). Deciphering complex protein interaction kinetics using interaction map. *Biochem. Biophys. Res. Commun.* **428** 74–79.
- ATHERTON, R., SCHAIKER, R. and DUCOT, E. (1975). On the statistical sensitivity analysis of models for chemical kinetics. *AIChE J.* **21** 441–448.
- BABACAN, S. D., MOLINA, R. and KATSAGGELOS, A. K. (2008). Parameter estimation in TV image restoration using variational distribution approximation. *IEEE Trans. Image Process.* **17** 326–339. [MR2492977](#)
- BERNARDO, J.-M. and SMITH, A. F. M. (1994). *Bayesian Theory*. *Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. Wiley, Chichester. [MR1274699](#)

Key words and phrases. Rate constant, biosensor, Bayesian, variational method, integral equation, adaptive discretization algorithm.

- BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning. Information Science and Statistics*. Springer, New York. [MR2247587](#)
- BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. [MR3671776](#)
- BOX, G. E. P. and DRAPER, N. R. (1965). The Bayesian estimation of common parameters from several responses. *Biometrika* **52** 355–365. [MR0210220](#)
- BROOKS, S., SMITH, J., VEHTARI, A., PLUMMER, M., STONE, M. et al. (2002). Discussion on the paper by Spiegelhalter, Best, Carlin and van der Linde. *J. Roy. Statist. Soc. Ser. B* **64** 616–639.
- CHEN, L., HOLST, M. and XU, J. (2009). Convergence and optimality of adaptive mixed finite element methods. *Math. Comp.* **78** 35–53. [MR2448696](#)
- EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. [MR2060166](#)
- ELMLUND, L., KÄCK, C., AASTRUP, T. and NICHOLLS, I. (2015). Study of the interaction of trastuzumab and SKOV3 epithelial cancer cells using a quartz crystal microbalance sensor. *Sensors* **15** 5884–5894.
- ENGL, H., HANKE, M. and NEUBAUER, A. (1996). *Regularization of Inverse Problems*. Kluwer, Dordrecht.
- EVANS, S. N. and STARK, P. B. (2002). Inverse problems as statistics. *Inverse Probl.* **18** R55–R97. [MR1929274](#)
- FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. [MR2530322](#)
- GORSHKOVA, I., SVITEL, J., RAZJOUYAN, F. and SCHUCK, P. (2008). Bayesian analysis of heterogeneity in the distribution of binding properties of immobilized surface sites. *Langmuir* **24** 11577–11586.
- GRAY, R. J. and PIERCE, D. A. (1985). Goodness-of-fit tests for censored survival data. *Ann. Statist.* **13** 552–563. [MR0790556](#)
- GUPTA, A. and RAWLINGS, J. (2014). Comparison of parameter estimation methods in stochastic chemical kinetic models: Examples in systems biology. *AIChE J.* **60** 1253–1268.
- JIN, B. and ZOU, J. (2010). Hierarchical Bayesian inference for ill-posed problems via variational method. *J. Comput. Phys.* **229** 7317–7343. [MR2677781](#)
- JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. and SAUL, L. (1999). Introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.
- KAPIO, J. and SOMERSALO, E. (2007). Statistical inverse problems: Discretization, model reduction and inverse crimes. *J. Comput. Appl. Math.* **198** 493–504. [MR2260683](#)
- KENNEDY, M. C. and O'HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. [MR1858398](#)
- KOMOROWSKI, M., COSTA, M., RAND, D. and STUMPF, M. (2011). Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proc. Natl. Acad. Sci. USA* **108** 8645–8650.
- KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, New York. [MR0103557](#)
- LOH, W.-Y. and ZHENG, W. (2013). Regression trees for longitudinal and multiresponse data. *Ann. Appl. Stat.* **7** 495–522. [MR3086428](#)
- PARISI, G. (1988). *Statistical Field Theory*. Addison-Wesley, New York.
- PETERSON, C. and ANDERSON, J. (1987). A mean field theory learning algorithm for neural networks. *Complex Systems* **1** 995–1019.
- PISCHEL, D., SUNDMACHER, K. and FLASSIG, R. (2017). Efficient simulation of intrinsic, extrinsic and external noise in biochemical systems. *Bioinformatics* **33** 1253–1268.
- RIVARA, M.-C. (1984). Mesh refinement processes based on the generalized bisection of simplices. *SIAM J. Numer. Anal.* **21** 604–613. [MR0744176](#)
- ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. [MR2080278](#)

- SANVICENS, N., MANNELLI, I., SALVADOR, J., VALERA, E. and MARCO, M. (2011). Biosensors for pharmaceuticals based on novel technology. *TrAC, Trends Anal. Chem.* **30** 541–553.
- STANLEY, B. and GUIOCHON, G. (1995). Calculation of adsorption energy distributions of silica samples using nonlinear chromatography. *Langmuir* **11** 1735–1743.
- STUART, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numer.* **19** 451–559. [MR2652785](#)
- SVITEL, J., BALBO, A., MARIUZZA, R., GONZALES, N. and SCHUCK, P. (2003). Combined affinity and rate constant distributions of ligand populations from experimental surface binding kinetics and equilibria. *Biophys. J.* **84** 4062–4077.
- TELESCA, D., MÜLLER, P., KORNBLAU, S. M., SUCHARD, M. A. and JI, Y. (2012). Modeling protein expression and protein signaling pathways. *J. Amer. Statist. Assoc.* **107** 1372–1384. [MR3036401](#)
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- TIKHONOV, A. N. and ARSENIN, V. Y. (1977). *Solutions of Ill-Posed Problems*. Winston & Sons, Washington, DC. [MR0455365](#)
- WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.
- WAND, M. P., ORMEROD, J. T., PADOAN, S. A. and FRÜHRWIRTH, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Anal.* **6** 847–900. [MR2869967](#)
- ZHANG, Y., FORSSÉN, P., FORNSTEDT, T., GULLIKSSON, M. and DAI, X. (2018). An adaptive regularization algorithm for recovering the rate constant distribution from biosensor data. *Inverse Probl. Sci. Eng.* **26** 1464–1489. [MR3827396](#)
- ZHANG, Y., YAO, Z., FORSSÉN, P. and FORNSTEDT, T. (2019). Supplement to “Estimating the rate constant from biosensor data via an adaptive variational Bayesian approach.” DOI:[10.1214/19-AOAS1263SUPP](#).
- ZIEGEL, E. and GORMAN, J. (1980). Kinetic modelling with multiresponse data. *Technometrics* **22** 139–151.
- ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the “degrees of freedom” of the lasso. *Ann. Statist.* **35** 2173–2192. [MR2363967](#)

ESTIMATING ABUNDANCE FROM MULTIPLE SAMPLING CAPTURE-RECAPTURE DATA VIA A MULTI-STATE MULTI-PERIOD STOPOVER MODEL

BY HANNAH WORTHINGTON^{*,1}, RACHEL MCCREA^{†,2}, RUTH KING[‡] AND
RICHARD GRIFFITHS[†]

University of St Andrews^{}, University of Kent[†] and University of Edinburgh[‡]*

Capture-recapture studies often involve collecting data on numerous capture occasions over a relatively short period of time. For many study species this process is repeated, for example, annually, resulting in capture information spanning multiple sampling periods. To account for the different temporal scales, the robust design class of models have traditionally been applied providing a framework in which to analyse all of the available capture data in a single likelihood expression. However, these models typically require strong constraints, either the assumption of closure within a sampling period (the closed robust design) or conditioning on the number of individuals captured within a sampling period (the open robust design). For real datasets these assumptions may not be appropriate. We develop a general modelling structure that requires neither assumption by explicitly modelling the movement of individuals into the population both within and between the sampling periods, which in turn permits the estimation of abundance within a single consistent framework. The flexibility of the novel model structure is further demonstrated by including the computationally challenging case of multi-state data where there is individual time-varying discrete covariate information. We derive an efficient likelihood expression for the new multi-state multi-period stopover model using the hidden Markov model framework. We demonstrate the significant improvement in parameter estimation using our new modelling approach in terms of both the multi-period and multi-state components through both a simulation study and a real dataset relating to the protected species of great crested newts, *Triturus cristatus*.

REFERENCES

- BESBEAS, P., FREEMAN, S. N., MORGAN, B. J. T. and CATCHPOLE, E. A. (2002). Integrating mark-recapture-recovery and census data to estimate animal abundance and demographic parameters. *Biometrics* **58** 540–547. [MR1933532](#)
- BROWNIE, C., HINES, J. E., NICHOLS, J. D., POLLOCK, K. H. and HESTBECK, J. B. (1993). Capture-recapture studies for multiple strata including non-Markovian transitions. *Biometrics* **49** 1173–1187.
- DUPUIS, J. A. and SCHWARZ, C. J. (2007). A Bayesian approach to the multistate Jolly–Seber capture-recapture model. *Biometrics* **63** 1015–1022, 1309. [MR2414577](#)

Key words and phrases. Hidden Markov model, great crested newts, multi-state data, individual time-varying discrete covariate.

- GIMENEZ, O., ROSSI, V., CHOQUET, R., DEHAIS, C., DORIS, B., VARELLA, H., VILA, J.-P. and PRADEL, R. (2007). State-space modelling of data on marked individuals. *Ecol. Model.* **206** 431–438.
- GRIFFITHS, R. A., FOSTER, J., WILKINSON, J. W. and SEWELL, D. (2015). Science, statistics and surveys: A herpetological perspective. *J. Appl. Ecol.* **52** 1413–1417.
- KENDALL, W. L. and BJORKLAND, R. (2001). Using open robust design models to estimate temporary emigration from capture-recapture data. *Biometrics* **57** 1113–1122. [MR1973818](#)
- KENDALL, W. L., POLLOCK, K. H. and BROWNIE, C. (1995). A likelihood-based approach to capture-recapture estimation of demographic parameters under the robust design. *Biometrics* **51** 293–308.
- KENDALL, W. L., STAPLETON, S., WHITE, G. C., RICHARDSON, J. I., PEARSON, K. N. and MASON, P. (2018). A multistate open robust design: Population dynamics, reproductive effort, and phenology of sea turtles from tagging data. *Ecol. Monogr.* **89** e1329.
- KING, R. (2012). A review of Bayesian state-space modelling of capture-recapture-recovery data. *Interface Focus* **2** 190–204.
- KING, R. (2014). Statistical ecology. *Annu. Rev. Stat. Appl.* **1** 401–426.
- KING, R. and LANGROCK, R. (2016). Semi-Markov Arnason–Schwarz models. *Biometrics* **72** 619–628. [MR3515788](#)
- KING, R. and MCCREA, R. S. (2014). A generalised likelihood framework for partially observed capture-recapture-recovery models. *Stat. Methodol.* **17** 30–45. [MR3133584](#)
- KING, R., MORGAN, B. J. T., GIMENEZ, O. and BROOKS, S. P. (2009). *Bayesian Analysis for Population Ecology*. CRC Press, Boca Raton, FL.
- LANGROCK, R. and KING, R. (2013). Maximum likelihood estimation of mark-recapture-recovery models in the presence of continuous covariates. *Ann. Appl. Stat.* **7** 1709–1732. [MR3127965](#)
- LEWIS, B., GRIFFITHS, R. A. and WILKINSON, J. W. (2017). Population status of great crested newts (*Triturus cristatus*) at sites subjected to development mitigation. *Herpetological Journal* **27** 133–142.
- MCCREA, R. S. and MORGAN, B. J. T. (2011). Multistate mark-recapture model selection using score tests. *Biometrics* **67** 234–241. [MR2898835](#)
- MCCREA, R. S., MORGAN, B. J. T., GIMENEZ, O., BESBEAS, P., LEBRETON, J.-D. and BREGNBALLE, T. (2010). Multi-site integrated population modelling. *J. Agric. Biol. Environ. Stat.* **15** 539–561. [MR2788639](#)
- PLEDGER, S., EFFORD, M., POLLOCK, K. H., COLLAZO, J. A. and LYONS, J. E. (2009). Stopover duration analysis with departure probability dependent on unknown time since arrival. In *Environmental and Ecological Statistics* **3** (D. L. Thomson, E. G. Cooch and M. J. Conroy, eds.) 349–363.
- POLLOCK, K. H. (1982). A capture-recapture design robust to unequal probability of capture. *The Journal of Wildlife Management* **46** 752–757.
- ROYLE, J. A. (2008). Modeling individual effects in the Cormack–Jolly–Seber model: A state-space formulation. *Biometrics* **64** 364–370, 664. [MR2432405](#)
- SCHOFIELD, M. R. and BARKER, R. J. (2008). A unified capture-recapture framework. *J. Agric. Biol. Environ. Stat.* **13** 458–477. [MR2590940](#)
- SCHWARZ, C. J. and ARNASON, A. N. (1996). A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics* **52** 860–873. [MR1411736](#)
- SCHWARZ, C. J. and STOBO, W. T. (1997). Estimating temporary migration using the robust design. *Biometrics* **53** 178–194.
- WORTHINGTON, H., MCCREA, R., KING, R. and GRIFFITHS, R. (2019). Supplement to “Estimating abundance from multiple sampling capture-recapture data via a multi-state multi-period stopover model.” DOI:10.1214/19-AOAS1264SUPPA, DOI:10.1214/19-AOAS1264SUPPB, DOI:10.1214/19-AOAS1264SUPPC.

ZUCCHINI, W., MACDONALD, I. L. and LANGROCK, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*, 2nd ed. *Monographs on Statistics and Applied Probability* **150**. CRC Press, Boca Raton, FL. [MR3618333](#)

ROBUST ELASTIC NET ESTIMATORS FOR VARIABLE SELECTION AND IDENTIFICATION OF PROTEOMIC BIOMARKERS

BY GABRIELA V. COHEN FREUE^{*,1}, DAVID KEPPLINGER^{*},
 MATÍAS SALIBIÁN-BARRERA^{*,1} AND EZEQUIEL SMUCLER^{*,†}

University of British Columbia^{} and Universidad Torcuato Di Tella[†]*

In large-scale quantitative proteomic studies, scientists measure the abundance of thousands of proteins from the human proteome in search of novel biomarkers for a given disease. Penalized regression estimators can be used to identify potential biomarkers among a large set of molecular features measured. Yet, the performance and statistical properties of these estimators depend on the loss and penalty functions used to define them. Motivated by a real plasma proteomic biomarkers study, we propose a new class of penalized robust estimators based on the elastic net penalty, which can be tuned to keep groups of correlated variables together in the selected model and maintain robustness against possible outliers. We also propose an efficient algorithm to compute our robust penalized estimators and derive a data-driven method to select the penalty term. Our robust penalized estimators have very good robustness properties and are also consistent under certain regularity conditions. Numerical results show that our robust estimators compare favorably to other robust penalized estimators. Using our proposed methodology for the analysis of the proteomics data, we identify new potentially relevant biomarkers of cardiac allograft vasculopathy that are not found with nonrobust alternatives. The selected model is validated in a new set of 52 test samples and achieves an area under the receiver operating characteristic (AUC) of 0.85.

REFERENCES

- ALFONS, A. (2016). robustHD: Robust Methods for High-Dimensional Data. R package version 0.5.1.
- ALFONS, A., CROUX, C. and GELPER, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Stat.* **7** 226–248. [MR3086417](#)
- CLARKE, F. H. (1990). *Optimization and Nonsmooth Analysis*, 2nd ed. *Classics in Applied Mathematics* **5**. SIAM, Philadelphia, PA. [MR1058436](#)
- COHEN FREUE, G. V. and BORCHERS, C. H. (2012). Multiple Reaction Monitoring (MRM). *Circ. Cardiovasc. Genet.* **5** 378.
- COHEN FREUE, G. V., KEPPLINGER, D., SALIBIÁN-BARRERA, M. and SMUCLER, E. (2019). Supplement to “Robust elastic net estimators for variable selection and identification of proteomic biomarkers.” DOI:[10.1214/19-AOAS1269SUPP](#).
- DOMANSKI, D., PERCY, A. J., YANG, J., CHAMBERS, A. G., HILL, J. S., FREUE, G. V. C. and BORCHERS, C. H. (2012). MRM-based multiplexed quantitation of 67 putative cardiovascular disease biomarkers in human plasma. *Proteomics* **12** 1222–1243.

Key words and phrases. Robust estimation, regularized estimation, penalized estimation, elastic net penalty, proteomics biomarkers.

- DONOHU, D. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann. Wadsworth Statist./Probab. Ser.* 157–184. Wadsworth, Belmont, CA. [MR0689745](#)
- EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. [MR2060166](#)
- FAN, J., LI, Q. and WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265. [MR3597972](#)
- FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. [MR2065194](#)
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.
- KHAN, J. A., VAN AELST, S. and ZAMAR, R. H. (2007). Robust linear model selection based on least angle regression. *J. Amer. Statist. Assoc.* **102** 1289–1299. [MR2412550](#)
- KOLLER, M. and STAHEL, W. A. (2017). Nonsingular subsampling for regression S estimators with categorical predictors. *Comput. Statist.* **32** 631–646. [MR3656977](#)
- LIN, D., FREUE, G. C., HOLLANDER, Z., MANCINI, G. B. J., SASAKI, M., MUI, A., WILSON-MCMANUS, J., IGNASZEWSKI, A., IMAI, C. et al. (2013). Plasma protein biosignatures for detection of cardiac allograft vasculopathy. *J. Heart Lung Transplant.* **32** 723–733.
- MARONNA, R. A. (2011). Robust ridge regression for high-dimensional data. *Technometrics* **53** 44–53. [MR2791951](#)
- MARONNA, R. A., MARTIN, R. D. and YOHAI, V. J. (2006). *Robust Statistics. Wiley Series in Probability and Statistics: Theory and Methods.* Wiley, Chichester. [MR2238141](#)
- MARONNA, R. A. and YOHAI, V. J. (2010). Correcting MM estimates for “fat” data sets. *Comput. Statist. Data Anal.* **54** 3168–3173. [MR2727743](#)
- OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000). On the LASSO and its dual. *J. Comput. Graph. Statist.* **9** 319–337. [MR1822089](#)
- PEÑA, D. and YOHAI, V. (1999). A fast procedure for outlier diagnostics in large regression problems. *J. Amer. Statist. Assoc.* **94** 434–445. [MR1702315](#)
- ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880. [MR0770281](#)
- ROUSSEEUW, P. and YOHAI, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis (Heidelberg, 1983). Lect. Notes Stat.* **26** 256–272. Springer, New York. [MR0786313](#)
- SALIBIAN-BARRERA, M. and YOHAI, V. J. (2006). A fast algorithm for S-regression estimates. *J. Comput. Graph. Statist.* **15** 414–427. [MR2246273](#)
- SCHMAUSS, D. and WEIS, M. (2008). Cardiac allograft vasculopathy. *Circ.* **117** 2131–2141.
- SMUCLER, E. and YOHAI, V. J. (2017). Robust and sparse estimators for linear regression models. *Comput. Statist. Data Anal.* **111** 116–130. [MR3630222](#)
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- TIBSHIRANI, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.* **7** 1456–1490. [MR3066375](#)
- TOMIOKA, R., SUZUKI, T. and SUGIYAMA, M. (2011). Super-linear convergence of dual augmented Lagrangian algorithm for sparsity regularized estimation. *J. Mach. Learn. Res.* **12** 1537–1586. [MR2813147](#)
- YOHAI, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656. [MR0888431](#)
- YOHAI, V. J. and ZAMAR, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *J. Amer. Statist. Assoc.* **83** 406–413. [MR0971366](#)
- 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](#)

ZOU, H. and ZHANG, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37** 1733–1751. [MR2533470](#)

STATISTICAL INFERENCE FOR PARTIALLY OBSERVED BRANCHING PROCESSES WITH APPLICATION TO CELL LINEAGE TRACKING OF *IN VIVO* HEMATOPOIESIS

BY JASON XU^{1,*}, SAMSON KOELLE^{3,†}, PETER GUTTORP[†],
 CHUANFENG WU^{3,‡}, CYNTHIA DUNBAR^{3,‡}, JANIS L. ABKOWITZ[†] AND
 VLADIMIR N. MININ^{2,§}

*Duke University**, *University of Washington†*, *National Institutes of Health‡* and
University of California, Irvine§

Single-cell lineage tracking strategies enabled by recent experimental technologies have produced significant insights into cell fate decisions, but lack the quantitative framework necessary for rigorous statistical analysis of mechanistic models describing cell division and differentiation. In this paper, we develop such a framework with corresponding moment-based parameter estimation techniques for continuous-time, multi-type branching processes. Such processes provide a probabilistic model of how cells divide and differentiate, and we apply our method to study *hematopoiesis*, the mechanism of blood cell production. We derive closed-form expressions for higher moments in a general class of such models. These analytical results allow us to efficiently estimate parameters of much richer statistical models of hematopoiesis than those used in previous statistical studies. To our knowledge, the method provides the first rate inference procedure for fitting such models to time series data generated from cellular barcoding experiments. After validating the methodology in simulation studies, we apply our estimator to hematopoietic lineage tracking data from rhesus macaques. Our analysis provides a more complete understanding of cell fate decisions during hematopoiesis in nonhuman primates, which may be more relevant to human biology and clinical strategies than previous findings from murine studies. For example, in addition to previously estimated hematopoietic stem cell self-renewal rate, we are able to estimate fate decision probabilities and to compare structurally distinct models of hematopoiesis using cross validation. These estimates of fate decision probabilities and our model selection results should help biologists compare competing hypotheses about how progenitor cells differentiate. The methodology is transferrable to a large class of stochastic compartmental and multi-type branching models, commonly used in studies of cancer progression, epidemiology and many other fields.

REFERENCES

- ABKOWITZ, J. L., LINENBERGER, M. L., NEWTON, M. A., SHELTON, G. H., OTT, R. L. and GUTTORP, P. (1990). Evidence for the maintenance of hematopoiesis in a large animal by the sequential activation of stem-cell clones. *Proc. Natl. Acad. Sci. USA* **87** 9062–9066.

Key words and phrases. Stochastic compartmental models, generalized method of moments, Markov jump processes, mechanistic modeling, stem cells.

- BAILEY, N. T. J. (1964). *The Elements of Stochastic Processes with Applications to the Natural Sciences*. Wiley, New York. [MR0165572](#)
- BECKER, A. J., MCCULLOCH, E. A. and TILL, J. E. (1963). Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature* **197** 452–454.
- BIASCO, L., PELLIN, D., SCALA, S., DIONISIO, F., BASSO-RICCI, L., LEONARDELLI, L., SCARAMUZZA, S., BARICORDI, C., FERRUA, F. et al. (2016). In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases. *Cell Stem Cell* **19** 107–119.
- BUCHHOLZ, V. R., FLOSSDORF, M., HENSEL, I., KRETSCHMER, L., WEISSBRICH, B., GRÄF, P., VERSCHOOR, A., SCHIEMANN, M., HÖFER, T. et al. (2013). Disparate individual fates compose robust CD8+ T cell immunity. *Science* **340** 630–635.
- CATLIN, S. N., ABKOWITZ, J. L. and GUTTORP, P. (2001). Statistical inference in a two-compartment model for hematopoiesis. *Biometrics* **57** 546–553. [MR1855690](#)
- CATLIN, S. N., BUSQUE, L., GALE, R. E., GUTTORP, P. and ABKOWITZ, J. L. (2011). The replication rate of human hematopoietic stem cells in vivo. *Blood* **117** 4460–4466.
- COLIJN, C. and MACKEY, M. C. (2005). A mathematical model of hematopoiesis. I. Periodic chronic myelogenous leukemia. *J. Theoret. Biol.* **237** 117–132. [MR2205754](#)
- DORMAN, K. S., SINSHEIMER, J. S. and LANGE, K. (2004). In the garden of branching processes. *SIAM Rev.* **46** 202–229. [MR2114452](#)
- FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. [MR1946581](#)
- FONG, Y., GUTTORP, P. and ABKOWITZ, J. (2009). Bayesian inference and model choice in a hidden stochastic two-compartment model of hematopoietic stem cell fate decisions. *Ann. Appl. Stat.* **3** 1695–1709. [MR2752154](#)
- GERRITS, A., DYKSTRA, B., KALMYKOWA, O. J., KLAUKE, K., VEROVSKAYA, E., BROEKHUIS, M. J. C., DE HAAN, G. and BYSTRYKH, L. V. (2010). Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood* **115** 2610–2618.
- GOLINELLI, D., GUTTORP, P. and ABKOWITZ, J. A. (2006). Bayesian inference in a hidden stochastic two-compartment model for feline hematopoiesis. *Math. Med. Biol.* **23** 153–172.
- GOYAL, S., KIM, S., CHEN, I. S. Y. and CHOU, T. (2015). Mechanisms of blood homeostasis: Lineage tracking and a neutral model of cell populations in rhesus macaques. *BMC Biol.* **13** 85.
- GRIFFIN, J. E. and BROWN, P. J. (2013). Some priors for sparse regression modelling. *Bayesian Anal.* **8** 691–702. [MR3102230](#)
- GUTTORP, P. (1995). *Stochastic Modeling of Scientific Data. Stochastic Modeling Series*. CRC Press, London. [MR1358359](#)
- HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. [MR0666123](#)
- HANSEN, L. P., HEATON, J. and YARON, A. (1996). Finite-sample properties of some alternative GMM estimators. *J. Bus. Econom. Statist.* **14** 262–280.
- HELLERSTEIN, M., HANLEY, M. B., CESAR, D., SILER, S., PAPAGEORGIOPOULOS, C., WIEDER, E., SCHMIDT, D., HOH, R., NEESE, R. et al. (1999). Directly measured kinetics of circulating T lymphocytes in normal and HIV-1-infected humans. *Nat. Med.* **5** 83–89.
- HYRIEN, O., PESLAK, S. A., YANEV, N. M. and PALIS, J. (2015). Stochastic modeling of stress erythropoiesis using a two-type age-dependent branching process with immigration. *J. Math. Biol.* **70** 1485–1521. [MR3343931](#)
- KAUR, A., DI MASCO, M., BARABASZ, A., ROSENZWEIG, M., MCCLURE, H. M., PERELSON, A. S., RIBEIRO, R. M. and JOHNSON, R. P. (2008). Dynamics of T- and B-lymphocyte turnover in a natural host of simian immunodeficiency virus. *J. Virol.* **82** 1084–1093.
- KAWAMOTO, H., WADA, H. and KATSURA, Y. (2010). A revised scheme for developmental pathways of hematopoietic cells: The myeloid-based model. *Int. Immunol.* **22** 65–70.

- KIMMEL, M. (2014). Stochasticity and determinism in models of hematopoiesis. In *A Systems Biology Approach to Blood* 119–152. Springer, New York.
- KIMMEL, M. and AXELROD, D. E. (2002). *Branching Processes in Biology. Interdisciplinary Applied Mathematics* **19**. Springer, New York. [MR1903571](#)
- KOELLE, S. J., ESPINOZA, D. A., WU, C., XU, J., LU, R., LI, B., DONAHUE, R. E. and DUNBAR, C. E. (2017). Quantitative stability of hematopoietic stem and progenitor cell clonal output in rhesus macaques receiving transplants. *Blood* **129** 1448–1457.
- LANGE, K. (2010). *Applied Probability*. Springer, New York.
- LASLO, P., PONGUBALA, J. M. R., LANCKI, D. W. and SINGH, H. (2008). Gene regulatory networks directing myeloid and lymphoid cell fates within the immune system. *Semin. Immunol.* **20** 228–235.
- LIEPE, J., KIRK, P., FILIPPI, S., TONI, T., BARNES, C. P. and STUMPF, M. P. H. (2014). A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat. Protoc.* **9** 439–456.
- LU, R., NEFF, N. F., QUAKE, S. R. and WEISSMAN, I. L. (2011). Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* **29** 928–933.
- MARCINIAK-CZOCRA, A., STIEHL, T., HO, A. D., JÄGER, W. and WAGNER, W. (2009). Modeling of asymmetric cell division in hematopoietic stem cells—regulation of self-renewal is essential for efficient repopulation. *Stem Cells Dev.* **18** 377–386.
- MARJORAM, P., MOLITOR, J., PLAGNOL, V. and TAVARÉ, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100** 15324–15328.
- NOTTA, F., ZANDI, S., TAKAYAMA, N., DOBSON, S., GAN, O. I., WILSON, G., KAUFMANN, K. B., MCLEOD, J., LAURENTI, E. et al. (2016). Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* aab2116.
- ORKIN, S. H. and ZON, L. I. (2008). Hematopoiesis: An evolving paradigm for stem cell biology. *Cell* **132** 631–644.
- PAKES, A. and POLLARD, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* **57** 1027–1057. [MR1014540](#)
- PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. [MR2524001](#)
- PERIÉ, L., HODGKIN, P., NAIK, S. H., SCHUMACHER, T. N., DE BOER, R. J. and DUFFY, K. R. (2014). Determining lineage pathways from cellular barcoding experiments. *Cell Rep.* **6** 617–624.
- PUDLO, P., MARIN, J. M., ESTOUP, A., CORNUET, J. M., GAUTIER, M. and ROBERT, C. P. (2016). Reliable ABC model choice via random forests. *Bioinformatics* **32** 859–866.
- SHEPHERD, B. E., KIEM, H. P., LANSDORP, P. M., DUNBAR, C. E., AUBERT, G., LAROCHELLE, A., SEGGEWISS, R., GUTTORP, P. and ABKOWITZ, J. L. (2007). Hematopoietic stem-cell behavior in nonhuman primates. *Blood* **110** 1806–1813.
- SIMINOVITCH, L., MCCULLOCH, E. A. and TILL, J. E. (1963). The distribution of colony-forming cells among spleen colonies. *J. Cell. Comp. Physiol.* **62** 327–336.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- TILL, J. E., MCCULLOCH, E. A. and SIMINOVITCH, L. (1964). A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. *Proc. Natl. Acad. Sci. USA* **51** 29–36.
- 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.
- VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. [MR1652247](#)

- VELTEN, L., HAAS, S. F., RAFFEL, S., BLASZKIEWICZ, S., ISLAM, S., HENNIG, B. P., HIRCHE, C., LUTZ, C., BUSS, E. C. et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19** 271.
- WAKEFIELD, J. (2013). *Bayesian and Frequentist Regression Methods. Springer Series in Statistics.* Springer, New York. [MR3025225](#)
- WEISSMAN, I. L. (2000). Stem cells: Units of development, units of regeneration, and units in evolution. *Cell* **100** 157–168.
- WHICHARD, Z. L., SARKAR, C. A., KIMMEL, M. and COREY, S. J. (2010). Hematopoiesis and its disorders: A systems biology approach. *Blood* **115** 2339–2347.
- WU, C., LI, B., LU, R., KOELLE, S. J., YANG, Y., JARES, A., KROUSE, A. E., METZGER, M., LIANG, F. et al. (2014). Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. *Cell Stem Cell* **14** 486–499.
- XU, J., KOELLE, S., GUTTORP, P., WU, C., DUNBAR, C., ABKOWITZ, J. L. and MININ, V. N. (2019). Supplement to “Statistical inference for partially observed branching processes with application to cell lineage tracking of *in vivo* hematopoiesis.” DOI:[10.1214/19-AOAS1272SUPP](#).
- ZHANG, Y., WALLACE, D. L., DE LARA, C. M., GHATTAS, H., ASQUITH, B., WORTH, A., GRIFFIN, G. E., TAYLOR, G. P., TOUGH, D. F. et al. (2007). In vivo kinetics of human natural killer cells: The effects of ageing and acute and chronic viral infection. *Immunology* **121** 258–265.

FIRE SEASONALITY IDENTIFICATION WITH MULTIMODALITY TESTS

BY JOSE AMEIJERAS-ALONSO^{*,1,2}, AKLI BENALI^{†,3},
ROSA M. CRUJEIRAS^{‡,1}, ALBERTO RODRÍGUEZ-CASAL^{†,1} AND
JOSÉ M. C. PEREIRA^{†,3}

*KU Leuven**, *Universidade de Lisboa[†]* and
Universidade de Santiago de Compostela[‡]

Understanding the role of vegetation fires in the Earth system is an important environmental problem. Although fire occurrence is influenced by natural factors, human activity related to land use and management has altered the temporal patterns of fire in several regions of the world. Hence, for a better insight into fires regimes it is of special interest to analyze where human activity has altered fire seasonality. For doing so, multimodality tests are a useful tool for determining the number of annual fire peaks. The periodicity of fires and their complex distributional features motivate the use of nonparametric circular statistics. The unsatisfactory performance of previous circular nonparametric proposals for testing multimodality justifies the introduction of a new approach, considering an adapted version of the excess mass statistic, jointly with a bootstrap calibration algorithm. A systematic application of the test on the Russia–Kazakhstan area is presented in order to determine how many fire peaks can be identified in this region. A False Discovery Rate correction, accounting for the spatial dependence of the data, is also required.

REFERENCES

- AMEIJEIRAS-ALONSO, J., CRUJEIRAS, R. M. and RODRÍGUEZ-CASAL, A. (2018). Applied directional statistics: Modern methods and case studies. In *Chap. Directional Statistics for Wildfires* (C. Ley and T. Verdebout, eds.) Chapman and Hall/CRC Press, Boca Raton, FL.
- AMEIJEIRAS-ALONSO, J., CRUJEIRAS, R. M. and RODRÍGUEZ-CASAL, A. (2019a). Mode testing, critical bandwidth and excess mass. *Test*. To appear. DOI:<https://www.doi.org/10.1007/s11749-018-0611-5>.
- AMEIJEIRAS-ALONSO, J., CRUJEIRAS, R. M. and RODRÍGUEZ-CASAL, A. (2019b). Supplement to “Fire seasonality identification with multimodality tests.” DOI:[10.1214/19-AOAS1273SUPP](https://doi.org/10.1214/19-AOAS1273SUPP).
- BENALI, A., MOTA, B., CARVALHAIS, N., OOM, D., MILLER, L. M., CAMPAGNOLO, M. L. and PEREIRA, J. M. C. (2017). Bimodal fire regimes unveil a global scale anthropogenic fingerprint. *Glob. Ecol. Biogeogr.* **26** 799–811.
- BENJAMINI, Y. and HELLER, R. (2007). False discovery rates for spatial signals. *J. Amer. Statist. Assoc.* **102** 1272–1281. [MR2412549](#)
- BENJAMINI, Y. and HOCHBERG, Y. (1997). Multiple hypotheses testing with weights. *Scand. J. Stat.* **24** 407–418. [MR1481424](#)
- BENJAMINI, Y., KRIEGER, A. M. and YEKUTIELI, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93** 491–507. [MR2261438](#)

Key words and phrases. Circular data, multimodality, multiple testing, wildfires.

- CHENG, M.-Y. and HALL, P. (1998). Calibrating the excess mass and dip tests of modality. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 579–589. [MR1625938](#)
- CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, New York. [MR1239641](#)
- FISHER, N. I. and MARRON, J. S. (2001). Mode testing via the excess mass estimate. *Biometrika* **88** 499–517. [MR1844848](#)
- GIGLIO, L., DESCLOITRES, J., JUSTICE, C. O. and KAUFMAN, Y. J. (2003). An enhanced contextual fire detection algorithm for MODIS. *Remote Sensing of Environment* **87** 273–282.
- HALL, P. and YORK, M. (2001). On the calibration of Silverman’s test for multimodality. *Statist. Sinica* **11** 515–536. [MR1844538](#)
- HALL, J. V., LOBODA, T. V., GIGLIO, L. and MCCARTY, G. W. (2016). A MODIS-based burned area assessment for Russian croplands: Mapping requirements and challenges. *Remote Sensing of Environment* **184** 506–521.
- HUCKEMANN, S., KIM, K.-R., MUNK, A., REHFELDT, F., SOMMERFELD, M., WEICKERT, J. and WOLLNIK, C. (2016). The circular SiZer, inferred persistence of shape parameters and application to early stem cell differentiation. *Bernoulli* **22** 2113–2142. [MR3498025](#)
- JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **186** 453–461. [MR0017504](#)
- KEELEY, J. E., SAFFORD, H., FOTHERINGHAM, C. J., FRANKLIN, J. and MORITZ, M. (2009). The 2007 southern California wildfires: Lessons in complexity. *Journal of Forestry* **107** 287–296.
- KORONTZI, S., MCCARTY, J., LOBODA, T., KUMAR, S. and JUSTICE, C. O. (2006). Global distribution of agricultural fires in croplands from 3 years of Moderate Resolution Imaging Spectroradiometer (MODIS) data. *Global Biogeochemical Cycles* **20** 202101–202115.
- LE PAGE, Y., OOM, D., SILVA, J., JÖNSSON, P. and PEREIRA, J. M. C. (2010). Seasonality of vegetation fires as modified by human action: Observing the deviation from eco-climatic fire regimes. *Glob. Ecol. Biogeogr.* **19** 575–588.
- LEY, C. and VERDEBOUT, T. (2017). *Modern Directional Statistics*. Chapman & Hall/CRC Interdisciplinary Statistics Series. CRC Press, Boca Raton, FL. [MR3752655](#)
- LOBODA, T. V., GIGLIO, L., BOSCHETTI, L. and JUSTICE, C. O. (2012). Regional fire monitoring and characterization using global NASA MODIS fire products in dry lands of Central Asia. *Front. Earth Sci.* **6** 196–205.
- MAGI, B. I., RABIN, S., SHEVLIAKOVA, E. and PACALA, S. (2012). Separating agricultural and non-agricultural fire seasonality at regional scales. *Biogeosciences* **9** 3003–3012.
- MÜLLER, D. W. and SAWITZKI, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86** 738–746. [MR1147099](#)
- NICHOLS, K., SCHOENBERG, F. P., KEELEY, J. E., BRAY, A. and DIEZ, D. (2011). The application of prototype point processes for the summary and description of California wildfires. *J. Time Series Anal.* **32** 420–429. [MR2857337](#)
- OLIVEIRA, M., CRUJEIRAS, R. M. and RODRÍGUEZ-CASAL, A. (2012). A plug-in rule for bandwidth selection in circular density estimation. *Comput. Statist. Data Anal.* **56** 3898–3908. [MR2957840](#)
- SILVERMAN, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43** 97–99. [MR0610384](#)
- WATSON, G. S. (1961). Goodness-of-fit tests on a circle. *Biometrika* **48** 109–114. [MR0131930](#)
- WESTERLING, A. L., GERSHUNOV, A., BROWN, T. J., CAYAN, D. R. and DETTINGER, M. D. (2003). Climate and wildfire in the western United States. *Bull. Am. Meteorol. Soc.* **84** 595–604.
- XU, H. and SCHOENBERG, F. P. (2011). Point process modeling of wildfire hazard in Los Angeles County, California. *Ann. Appl. Stat.* **5** 684–704. [MR2840171](#)

JOINT MODEL OF ACCELERATED FAILURE TIME AND MECHANISTIC NONLINEAR MODEL FOR CENSORED COVARIATES, WITH APPLICATION IN HIV/AIDS¹

BY HONGBIN ZHANG AND LANG WU

City University of New York and University of British Columbia

For a time-to-event outcome with censored time-varying covariates, a joint Cox model with a linear mixed effects model is the standard modeling approach. In some applications such as AIDS studies, *mechanistic* nonlinear models are available for some covariate process such as viral load during anti-HIV treatments, derived from the underlying data-generation mechanisms and disease progression. Such a mechanistic nonlinear covariate model may provide better-predicted values when the covariates are left censored or mismeasured. When the focus is on the impact of the time-varying covariate process on the survival outcome, an accelerated failure time (AFT) model provides an excellent alternative to the Cox proportional hazard model since an AFT model is formulated to allow the influence of the outcome by the entire covariate process. In this article, we consider a nonlinear mixed effects model for the censored covariates in an AFT model, implemented using a Monte Carlo EM algorithm, under the framework of a joint model for simultaneous inference. We apply the joint model to an HIV/AIDS data to gain insights for assessing the association between viral load and immunological restoration during antiretroviral therapy. Simulation is conducted to compare model performance when the covariate model and the survival model are misspecified.

REFERENCES

- BERNHARDT, P. W., WANG, H. J. and ZHANG, D. (2014). Flexible modeling of survival data with covariates subject to detection limits via multiple imputation. *Comput. Statist. Data Anal.* **69** 81–91. [MR3146878](#)
- BOOTH, J. G. and HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. Roy. Statist. Soc. Ser. B* **61** 265–285.
- BRESLAW, J. A. (1994). Random sampling from a truncated multivariate normal distribution. *Appl. Math. Lett.* **7** 1–6. [MR1349883](#)
- COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. [MR0341758](#)
- COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data. Monographs on Statistics and Applied Probability*. Chapman & Hall, London. [MR0751780](#)
- DEEKS, S., BARBOUR, J., MARTIN, J. et al. (2000). Sustained CD4+ T cell response after virologic failure of protease inhibitor-based regimens in patients with human immunodeficiency virus infection. *J. Infect. Dis.* **181** 946–953.

Key words and phrases. Censored data, HIV/AIDS, mechanistic model, nonlinear mixed effects model, survival data.

- FISCHL, M., RIBAUDO, H. and COLLIER, A. (2003). A randomized trial of 2 different 4-drug antiretroviral regimens versus a 3-drug regimen, in advanced human immunodeficiency virus disease. *J. Infect. Dis.* **188** 625–634.
- GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. [MR1141740](#)
- GEWEKE, J. (1996). *Handbook of Computational Economics*. North-Holland, Amsterdam.
- GILKS, W. and WILD, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41** 337–348.
- HUANG, Y., HU, X. J. and DAGNE, G. A. (2014). Jointly modeling time-to-event and longitudinal data: A Bayesian approach. *Stat. Methods Appl.* **23** 95–121. [MR3171402](#)
- IBRAHIM, J. G., LIPSITZ, S. R. and CHEN, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 173–190. [MR1664045](#)
- KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, New York. [MR1924807](#)
- LEDERMAN, M. M., CONNICK, E. and LANDY, A. (1998). Immunologic responses associated with 12 weeks of combination antiretroviral therapy consisting of zidovudine, lamivudine, and zalcitabine: Results of AIDS clinical trials group protocol 315. *Journal of Infectious Disease* **178** 79–79.
- LINDSTROM, M. J. and BATES, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46** 673–687. [MR1085815](#)
- MCLACHLAN, G. J. and KRISHNAN, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York. [MR1417721](#)
- MILLER, R. G. (1976). *Survival Analysis*, 2nd ed. Wiley Classics Library, New York.
- PALMISANO, L. and YELLA, L. (2011). A brief history of antiretroviral therapy of HIV infection: Success and challenges. *Ann. Ist. Super. Sanita* **47** 44–48.
- PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. and FLANNERY, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. Cambridge Univ. Press, New York.
- RIOPOULOS, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data*. Chapman & Hall, London.
- SATTAR, A. and SINHA, S. K. (2019). Joint modeling of longitudinal and survival data with a covariate subject to a limit of detection. *Stat. Methods Med. Res.* **28** 486–502. [MR3903754](#)
- TSENG, Y.-K., HSIEH, F. and WANG, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92** 587–603. [MR2202648](#)
- TSENG, Y. K. and YANG, Y. F. (2016). A kernel smooth approach for joint modeling of accelerated failure time and longitudinal data. *Comm. Statist. Simulation Comput.* **45** 1240–1248. [MR3486657](#)
- VAIDA, F. and LIU, L. (2009). Fast implementation for normal mixed effects models with censored response. *J. Comput. Graph. Statist.* **18** 797–817. [MR2750442](#)
- WEI, G. C. and TANNER, M. A. (1990). A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. Amer. Statist. Assoc.* **85** 699–704.
- WU, L. (2010). *Mixed Effects Models for Complex Data*. Chapman & Hall, London.
- WU, H. and DING, A. (1999). Population HIV-1 dynamics in vivo: Applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics* **55** 410–418.
- WU, L., LIU, W. and HU, X. J. (2010). Joint inference on HIV viral dynamics and immune suppression in presence of measurement errors. *Biometrics* **66** 327–335. [MR2758812](#)
- WU, H., ZHAO, C. and LIANG, H. (2004). Comparison of linear, nonlinear and semiparametric mixed-effects models for estimating HIV dynamic parameters. *Biom. J.* **46** 233–245. [MR2099362](#)
- ZHANG, H., WONG, H. and WU, L. (2018). A mechanistic nonlinear model for censored and mis-measured covariates in longitudinal models, with application in AIDS studies. *Stat. Med.* **37** 167–178. [MR3738069](#)

ZHANG, H. and WU, L. (2018). A non-linear model for censored and mismeasured time varying covariates in survival models, with applications in human immunodeficiency virus and acquired immune deficiency syndrome studies. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 1437–1450. [MR3873715](#)

PREDICTION OF SMALL AREA QUANTILES FOR THE CONSERVATION EFFECTS ASSESSMENT PROJECT USING A MIXED EFFECTS QUANTILE REGRESSION MODEL¹

BY EMILY BERG AND DANHYANG LEE

Iowa State University and University of Alabama

Quantiles of the distributions of several measures of erosion are important parameters in the Conservation Effects Assessment Project, a survey intended to quantify soil and nutrient loss on crop fields. Because sample sizes for domains of interest are too small to support reliable direct estimators, model based methods are needed. Quantile regression is appealing for CEAP because finding a single family of parametric models that adequately describes the distributions of all variables is difficult and small area quantiles are parameters of interest. We construct empirical Bayes predictors and bootstrap mean squared error estimators based on the linearly interpolated generalized Pareto distribution (LIGPD). We apply the procedures to predict county-level quantiles for four types of erosion in Wisconsin and validate the procedures through simulation.

REFERENCES

- BATTESE, G. E., HARTER, R. and FULLER, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.* **83** 28–36.
- BERG, E. and CHANDRA, H. (2014). Small area prediction for a unit-level lognormal model. *Comput. Statist. Data Anal.* **78** 159–175. [MR3212164](#)
- BERG, E. and LEE, D. (2019). Supplement to “Prediction of small area quantiles for the conservation effects assessment project using a mixed effects quantile regression model.” DOI:[10.1214/19-AOAS1276SUPP](https://doi.org/10.1214/19-AOAS1276SUPP).
- CDC. <https://wonder.cdc.gov/nasa-nldas.html>.
- CHAMBERS, R. and TZAVIDIS, N. (2006). *M*-Quantile models for small area estimation. *Biometrika* **93** 255–268. [MR2278081](#)
- CHEN, J. and LIU, Y. (2012). Small area estimation under density ratio model. In *JSM Proceedings* 5162–5173. Amer. Statist. Assoc., Alexandria, VA.
- CHEN, J. and LIU, Y. (2017). Small area quantile estimation. Available at [arXiv:1705.10063](https://arxiv.org/abs/1705.10063).
- CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and GALICHON, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika* **96** 559–575. [MR2538757](#)
- DIALLO, M. S. and RAO, J. N. K. (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scand. J. Stat.* **45** 1092–1116. [MR3884901](#)
- FENG, Y., CHEN, Y. and HE, X. (2015). Bayesian quantile regression with approximate likelihood. *Bernoulli* **21** 832–850. [MR3338648](#)
- GERACI, M. and BOTTAI, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8** 140–154.

Key words and phrases. Quantile regression, empirical Bayes, parametric bootstrap, erosion, environmental monitoring.

- GERACI, M. and BOTTAI, M. (2014). Linear quantile mixed models. *Stat. Comput.* **24** 461–479. [MR3192268](#)
- GERACI, M. and JONES, M. C. (2015). Improved transformation-based quantile regression. *Canad. J. Statist.* **43** 118–132. [MR3324431](#)
- GOEBEL, J. J. (2009). Statistical methodology for the NRI-CEAP cropland survey. USDA/NRCS.
- GOEBEL, J. J. and KELLOGG, R. L. (2002). Using survey data and modeling to assist the development of agri-environmental policy. In *Conference on Agricultural and Environmental Statistical Applications in Rome* 695–705. National Statistical Institute of Italy, Rome.
- HYNDMAN, R. J. and FAN, Y. (1996). Sample quantiles in statistical packages. *Amer. Statist.* **50** 361–365.
- JANG, W. and WANG, H. J. (2015). A semiparametric Bayesian approach for joint-quantile regression with clustered data. *Comput. Statist. Data Anal.* **84** 99–115. [MR3292800](#)
- JIANG, J. and LAHIRI, P. (2006). Mixed model prediction and small area estimation. *TEST* **15** 1–96. [MR2252522](#)
- KOENKER, R. (2005). *Quantile Regression. Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. [MR2268657](#)
- KOENKER, R. and NG, P. (2005). Inequality constrained quantile regression. *Sankhyā* **67** 418–440. [MR2208897](#)
- MOLINA, I., NANDRAM, B. and RAO, J. N. K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *Ann. Appl. Stat.* **8** 852–885. [MR3262537](#)
- MOLINA, I. and RAO, J. N. K. (2010). Small area estimation of poverty indicators. *Canad. J. Statist.* **38** 369–385. [MR2730115](#)
- NUSSER, S. M. and GOEBEL, J. J. (1997). The national resources inventory: A long-term multi-resource monitoring programme. *Environ. Ecol. Stat.* **4** 181–204.
- PFEFFERMANN, D. (2013). New important developments in small area estimation. *Statist. Sci.* **28** 40–68. [MR3075338](#)
- PORTNOY, S. (1991). Asymptotic behavior of the number of regression quantile breakpoints. *SIAM J. Sci. Statist. Comput.* **12** 867–883. [MR1102413](#)
- RAO, J. N. K. and MOLINA, I. (2015). *Small Area Estimation*, 2nd ed. *Wiley Series in Survey Methodology*. Wiley, Hoboken, NJ. [MR3380626](#)
- REICH, B. J., FUENTES, M. and DUNSON, D. B. (2011). Bayesian spatial quantile regression. *J. Amer. Statist. Assoc.* **106** 6–20. [MR2816698](#)
- USDA (2015). Summary report: 2012 National Resources Inventory, Natural Resources Conservation Service, Washington, DC, and Center for Survey Statistics and Methodology, Iowa State Univ., Ames, IA.
- USDA/NRCS (2012). Assessment of the effects of conservation practices on cultivated cropland in the upper Mississippi river basin. Conservation Effects Assessment Project. U.S. Dept. Agriculture. Available at https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1042093.pdf.
- WANG, J., FULLER, W. A. and QU, Y. (2008). Small area estimation under a restriction. *Surv. Methodol.* **34**, 29–36.
- WEIDENHAMMER, B., SCHMID, T., SALVATI, N. and TZAVIDIS, N. (2016). A unit-level quantile nested error regression model for domain prediction with continuous and discrete outcomes. Economics Discussion Paper, School of Business and Economics, Freie Univ., Berlin.
- WILLIAMS, J. R. and IZAURRALDE, R. C. (2006). The APEX model. In *Watershed Models* (V. P. Singh and D. K. Frevert, eds.) 437–482. CRC Press, Boca Raton, FL.
- WISCHMEIER, W. H. and SMITH, D. D. (1978). Predicting rainfall erosion losses—a guide to conservation planning. U.S. Dept. Agriculture Handbook No. 537.
- WOODRUFF, R. S. (1952). Confidence intervals for medians and other position measures. *J. Amer. Statist. Assoc.* **47** 635–646. [MR0050845](#)

MICROSIMULATION MODEL CALIBRATION USING INCREMENTAL MIXTURE APPROXIMATE BAYESIAN COMPUTATION

BY CAROLYN M. RUTTER^{*,1}, JONATHAN OZIK^{†,‡,2}, MARIA DEYOREO^{*,1}
 AND NICHOLSON COLLIER^{†,‡,2}

RAND Corporation^{}, University of Chicago[†] and Argonne National Laboratory[‡]*

Microsimulation models (MSMs) are used to inform policy by predicting population-level outcomes under different scenarios. MSMs simulate individual-level event histories that mark the disease process (such as the development of cancer) and the effect of policy actions (such as screening) on these events. MSMs often have many unknown parameters; calibration is the process of searching the parameter space to select parameters that result in accurate MSM prediction of a wide range of targets. We develop Incremental Mixture Approximate Bayesian Computation (IMABC) for MSM calibration which results in a simulated sample from the posterior distribution of model parameters given calibration targets. IMABC begins with a rejection-based ABC step, drawing a sample of points from the prior distribution of model parameters and accepting points that result in simulated targets that are near observed targets. Next, the sample is iteratively updated by drawing additional points from a mixture of multivariate normal distributions and accepting points that result in accurate predictions. Posterior estimates are obtained by weighting the final set of accepted points to account for the adaptive sampling scheme. We demonstrate IMABC by calibrating CRC-SPIN 2.0, an updated version of a MSM for colorectal cancer (CRC) that has been used to inform national CRC screening guidelines.

REFERENCES

- BEAUMONT, M. A., CORNUET, J.-M., MARIN, J.-M. and ROBERT, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika* **96** 983–990. [MR2767283](#)
- BLATT, L. J. (1961). Polyps of the colon and rectum: Incidence and distribution. *Dis. Colon Rectum* **4** 277–282.
- BLUM, M. G. B. and FRANÇOIS, O. (2010). Non-linear regression models for approximate Bayesian computation. *Stat. Comput.* **20** 63–73. [MR2578077](#)
- BOMBI, J. A. (1988). Polyps of the colon in Barcelona, Spain. *Cancer* **61** 1472–1476.
- CHAPMAN, I. (1963). Adenomatous polypi of large intestine: Incidence and distribution. *Ann. Surg.* **157** 223–226.
- CHEN, T. H. H., YEN, M. F., LAI, M. S., KOONG, S. L., WANG, C. Y., WONG, J. M., PREVOST, T. C. and DUFFY, S. W. (1999). Evaluation of a selective screening for colorectal carcinoma: The Taiwan multicenter cancer screening (TAMACS) project. *Cancer* **86** 1116–1128.
- CHURCH, J. M. (2004). Clinical significance of small colorectal polyps. *Dis. Colon Rectum* **47** 481–485.

Key words and phrases. Adaptive ABC, agent-based models, colorectal cancer.

- CONLAN, A. J. K., MCKINLEY, T. J., KAROLEMEAS, K., POLLOCK, E. B., GOODCHILD, A. V., MITCHELL, A. P., BIRCH, C. P. D., CLIFTON-HADLEY, R. S. and WOOD, J. L. N. (2012). Estimating the hidden burden of bovine tuberculosis in Great Britain. *PLoS Comput. Biol.* **8** e1002730. [MR3005930](#)
- CORLEY, D. A., JENSEN, C. D., MARKS, A. R., ZHAO, W. K., DE BOER, J., LEVIN, T. R., DOUBENI, C., FIREMAN, B. H. and QUESENBERRY, C. P. (2013). Variation of adenoma prevalence by age, sex, race, and colon location in a large population: Implications for screening and quality programs. *Clin. Gastroenterol. Hepatol.* **11** 172–180.
- DE KONING, H. J., MEZA, R., PLEVITIS, S. K., TEN HAAF, K., MUNSHI, V. N., JEON, J., ERDOGAN, S. A., KONG, C. Y., HAN, S. S. et al. (2014). Benefits and harms of computed tomography lung cancer screening strategies: A comparative modeling study for the US Preventive Services Task Force. *Ann. Intern. Med.* **160** 311–320.
- EIDE, T. J. and STALSBERG, H. (1978). Polyps of the large intestine in Northern Norway. *Cancer* **42** 2839–2848.
- CENTERS FOR DISEASE CONTROL (2011). Vital signs: Colorectal cancer screening, incidence, and mortality—United States, 2002–2010. *Morb. Mortal. Wkly. Rep.* **60** 884–889.
- FRAZIER, D. T., MARTIN, G. M., ROBERT, C. P. and ROUSSEAU, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika* **105** 593–607. [MR3842887](#)
- HAKULINEN, T. (1977). On long-term relative survival rates. *J. Clin. Epidemiol.* **30** 431–443.
- HIXSON, L., FENNERTY, M., SAMPLINER, R., MCGEE, D. and GAREWAL, H. (1990). Prospective study of the frequency and size distribution of polyps missed by colonoscopy. *J. Natl. Cancer Inst.* **82** 1769–1772.
- IMPERIALE, T. F., WAGNER, D. R., LIN, C. Y., LARKIN, G. N., ROGGE, J. D. and RANSOHOFF, D. F. (2000). Risk of advanced proximal neoplasms in asymptomatic adults according to the distal colorectal findings. *N. Engl. J. Med.* **343** 169–174.
- JOHANNSEN, L. G. K., MOMSEN, O. and JACOBSEN, N. O. (1989). Polyps of the large intestine in Aarhus, Denmark. An autopsy study. *Cancer* **24** 799–806.
- KIM, J. J., BURGER, E. A., REGAN, C. and SY, S. (2017). Screening for cervical cancer in primary care: a decision analysis for the U.S. Preventive Services Task Force. Technical Report. Agency for Healthcare Research and Quality. Available at <https://www.uspreventiveservicestaskforce.org/Home/GetFileByID/3287>, Contract No. HHSA-290-2012-00015-I. Last accessed on February 1, 2019.
- KISH, L. (1965). *Survey Sampling*. Wiley, New York, USA.
- KNUDSEN, A. B., ZAUBER, A. G., RUTTER, C. M., NABER, S. K., DORIA-ROSE, V. P., PABINIAK, C., JOHANSON, C., FISCHER, S. E., LANSDORP-VOGELAAR, I. et al. (2016). Estimation of benefits, burden, and harms of colorectal cancer screening strategies: Modeling study for the US Preventive Services Task Force. *J. Am. Med. Dir. Assoc.* **315** 2595–2609.
- KOH, K.-J., LIN, L.-H., HUANG, S.-H. and WONG, J.-U. (2015). CARE—pediatric colon adenocarcinoma: A case report and literature review comparing differences in clinical features between children and adult patients. *Medicine (Baltim. Md.)* **94** e503.
- KONG, C. Y., MCMAHON, P. M. and GAZELLE, G. S. (2009). Calibration of disease simulation model using an engineering approach. *Value Health* **12** 521–529.
- LESLIE, A., CAREY, F. A., PRATT, N. R. and STEELE, R. J. C. (2002). The colorectal adenoma-carcinoma sequence. *Br. J. Surg.* **89** 845–860.
- LI, W. and FEARNHEAD, P. (2018). On the asymptotic efficiency of approximate Bayesian computation estimators. *Biometrika* **105** 285–299. [MR3804403](#)
- LIEBERMAN, D., MORAVEC, M., HOLUB, J., MICHAELS, L. and EISEN, G. (2008). Polyp size and advanced histology in patients undergoing colonoscopy screening: Implications for CT colonography. *Gastroenterology* **135** 1100–1105.
- LIU, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer, New York. [MR1842342](#)

- MANDELBLATT, J. S., STOUT, N. K., SCHECHTER, C. B., VAN DEN BROEK, J. J., MIGLIORETTI, D. L., KRAPCHO, M., TRENTAM-DIETZ, A., MUNOZ, D., LEE, S. J. et al. (2016). Collaborative modeling of the benefits and harms associated with different US breast cancer screening strategies. *Ann. Intern. Med.* **164** 215–225.
- 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 SIMON, T. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100** 15324–15328.
- MCKINLEY, T. J., VERNON, I., ANDRIANAKIS, I., MCCREESH, N., OAKLEY, J. E., NSUBUGA, R. N., GOLDSTEIN, M. and WHITE, R. G. (2018). Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models. *Statist. Sci.* **33** 4–18. [MR3757500](#)
- MEISSNER, H. I., BREEN, N., KLABUNDE, C. N. and VERNON, S. W. (2006). Patterns of colorectal cancer screening uptake among men and women in the United States. *Cancer Epidemiol. Biomark. Prev.* **15** 389–394.
- MUTO, T., BUSSEY, H. J. R. and MORSON, B. C. (1975). The evolution of cancer in the colon and rectum. *Cancer* **36** 2251–2270.
- NATIONAL CANCER INSTITUTE (2004). Surveillance, epidemiology, and end results (SEER) program. Available at <http://www.seer.cancer.gov>, SEER Stat Database: Incidence—SEER 9 Regs Public-Use, Nov 2003 Sub (1973–2001), released April 2004, based on the November 2003 submission.
- NATIONAL CANCER INSTITUTE (2018). Cancer INtervention and Surveillance Modeling Network (CISNET). Available at <https://cisnet.cancer.gov>. Last accessed on February 1, 2019.
- NATIONAL CENTER FOR HEALTH STATISTICS (2000). US Life Tables.
- NELDER, J. A. and MEAD, R. (1965). A simplex method for function minimization. *Comput. J.* **7** 308–313. [MR3363409](#)
- OZIK, J., COLLIER, N. T., WOZNIAK, J. M. and SPAGNUOLO, C. (2016). From desktop to large-scale model exploration with Swift/T. In *Proceedings of the 2016 Winter Simulation Conference (WSC)* 206–220. IEEE Press.
- PICKHARDT, P. J., CHOI, R., HWANG, I., BUTLER, J. A., PUCKETT, M. L., HILDEBRANDT, H. A., WONG, R. K., NUGENT, P. A., MYSLIWIEC, P. A. et al. (2003). Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults. *N. Engl. J. Med.* **349** 2191–2200.
- PONUGOTI, P. L. and REX, D. K. (2017). Yield of a second screening colonoscopy 10 years after an initial negative examination in average-risk individuals. *Gastroint. Endosc.* **85** 221–224.
- PRITCHARD, J. K., SEIELSTAD, M. T., PEREZ-LEZAUN, A. and FELDMAN, M. W. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16** 1791–1798.
- R CORE TEAM (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at <http://www.R-project.org/>. Last accessed on February 1, 2019.
- RAFTERY, A. E. and BAO, L. (2010). Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. *Biometrics* **66** 1162–1173. [MR2758504](#)
- RATMANN, O., CAMACHO, A., MEIJER, A. and DONKER, G. (2014). Statistical modelling of summary values leads to accurate approximate Bayesian computations. Available at [arXiv:1305.4283](https://arxiv.org/abs/1305.4283).
- REX, D., CUTLER, C., LEMMEL, G., RAHMANI, E., CLARK, D., HELPER, D., LEHMAN, G. and MARK, D. (1997). Colonoscopic miss rates of adenomas determined by back-to-back colonoscopies. *Gastroenterology* **112** 24–28.
- RICKERT, R. R., AUERBACH, O., GARFINKEL, L., HAMMOND, E. C. and FRASCA, J. M. (1979). Adenomatous lesions of the large bowel. An autopsy survey. *Cancer* **43** 1847–1857.

- RUBIN, D. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *J. Amer. Statist. Assoc.* **82** 543–546.
- RUTTER, C. M., MIGLIORETTI, D. L. and SAVARINO, J. E. (2009). Bayesian calibration of microsimulation models. *J. Amer. Statist. Assoc.* **104** 1338–1350. [MR2750568](#)
- RUTTER, C. M. and SAVARINO, J. E. (2010). An evidence-based microsimulation model for colorectal cancer: Validation and application. *Cancer Epidemiol. Biomark. Prev.* 1055–9965.
- RUTTER, C. M., JOHNSON, E. A., FEUER, E. J., KNUDSEN, A. B., KUNTZ, K. M. and SCHRAG, D. (2013). Secular trends in colon and rectal cancer relative survival. *J. Natl. Cancer Inst.* **105** 1806–1813.
- RUTTER, C. M., KNUDSEN, A. B., MARSH, T. L., DORIA-ROSE, V. P., JOHNSON, E., PABINIAK, C., KUNTZ, K. M., VAN BALLEGOOIJEN, M., ZAUBER, A. G. et al. (2016). Validation of models used to inform colorectal cancer screening guidelines: Accuracy and implications. *Med. Decis. Mak.* **36** 604–614.
- SISSON, S. A., FAN, Y. and BEAUMONT, M. A. (2019). Overview of ABC. In *Handbook of Approximate Bayesian Computation*. Chapman & Hall/CRC Handb. Mod. Stat. Methods 3–54. CRC Press, Boca Raton, FL. [MR3889278](#)
- SISSON, S. A., FAN, Y. and TANAKA, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **104** 1760–1765. [MR2301870](#)
- STEELE, R. J., RAFTERY, A. E. and EMOND, M. J. (2006). Computing normalizing constants for finite mixture models via incremental mixture importance sampling (IMIS). *J. Comput. Graph. Statist.* **15** 712–734. [MR2291269](#)
- STEMMERMANN, G. N. and YATANI, R. (1973). Diverticulosis and polyps of the large intestine. A necropsy study of Hawaii Japanese. *Cancer* **31** 1260–1270.
- SZCZEPANSKI, W., URBAN, A. and WIERZCHOWSKI, W. (1992). Colorectal polyps in autopsy material. Part I. Adenomatous polyps. *Patol. Pol.* **43** 79–85.
- TAVARE, S., BALDING, D., GRIFFITHS, R. and DONNELLY, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145** 505–518.
- TJØRVE, E. and TJØRVE, K. M. (2010). A unified approach to the Richards-model family for use in growth analyses: Why we need only two model forms. *J. Theoret. Biol.* **267** 417–425.
- TONI, T., WELCH, D., STRELKOWA, N. and STUMPF, M. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6** 187–202.
- WILLIAMS, A. R., BALASOORIYA, B. A. W. and DAY, D. W. (1982). Polyps and cancer of the large bowel: A necropsy study in Liverpool. *Gut* **23** 835–842.
- WINAWER, S. J., FLETCHER, R. H., MILLER, L., GODLEE, F., STOLAR, M., MULROW, C., WOOLF, S., GLICK, S., GANIATS, T. et al. (1997). Colorectal cancer screening: Clinical guidelines and rationale. *Gastroenterology* **112** 594–642.
- WOZNIAK, J. M., ARMSTRONG, T. G., WILDE, M., KATZ, D. S., LUSK, E. and FOSTER, I. T. (2013). Swift/T: Large-scale application composition via distributed-memory dataflow processing. In *2013 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing* 95–102. IEEE.
- ZAUBER, A. G., KNUDSEN, A. B., RUTTER, C. M., LANSDORP-VOGELAAR, I., SAVARINO, J. E., VAN BALLEGOOIJEN, M. and KUNTZ, K. M. (2009). Cost-Effectiveness of CT colonography to screen for colorectal cancer: Report to the Agency for Healthcare Research and Quality from the Cancer Intervention and Surveillance Modeling Network (CISNET) for MIS-CAN, SimCRC, and CRC-SPIN Models. Technical Report. Available at <https://www.cms.gov/medicare-coverage-database/details/technology-assessments-details.aspx?TAId=58>, Project ID: CTCC0608, Last accessed on February 1, 2019.

PRINCIPAL NESTED SHAPE SPACE ANALYSIS OF MOLECULAR DYNAMICS DATA¹

BY IAN L. DRYDEN^{*}, KWANG-RAE KIM[†], CHARLES A. LAUGHTON^{*} AND
 HUILING LE^{*}

University of Nottingham^{} and SAS Korea[†]*

Molecular dynamics simulations produce huge datasets of temporal sequences of molecules. It is of interest to summarize the shape evolution of the molecules in a succinct, low-dimensional representation. However, Euclidean techniques such as principal components analysis (PCA) can be problematic as the data may lie far from in a flat manifold. Principal nested spheres gives a fundamentally different decomposition of data from the usual Euclidean subspace based PCA [*Biometrika* **99** (2012) 551–568]. Subspaces of successively lower dimension are fitted to the data in a backwards manner with the aim of retaining signal and dispensing with noise at each stage. We adapt the methodology to 3D subshape spaces and provide some practical fitting algorithms. The methodology is applied to cluster analysis of peptides, where different states of the molecules can be identified. Also, the temporal transitions between cluster states are explored.

REFERENCES

- COOTES, T. F., TAYLOR, C. J., COOPER, D. H. and GRAHAM, J. (1994). Image search using flexible shape models generated from sets of examples. In *Statistics and Images, Vol. 2* (K. V. Mardia, ed.) 111–139. Carfax, Oxford.
- DAMON, J. and MARRON, J. S. (2014). Backwards principal component analysis and principal nested relations. *J. Math. Imaging Vision* **50** 107–114. [MR3233137](#)
- DRYDEN, I. L. (2018). *shapes* package. R Foundation for Statistical Computing, Vienna, Austria. Contributed package, Version 1.2.4.
- DRYDEN, I. L. and MARDIA, K. V. (2016). *Statistical Shape Analysis, with Applications in R*, 2nd ed. Wiley, Chichester.
- FLETCHER, P. T., LU, C., PIZER, S. and JOSHI, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. Med. Imag.* **23** 995–1005.
- GOODALL, C. (1991). Procrustes methods in the statistical analysis of shape. *J. Roy. Statist. Soc. Ser. B* **53** 285–339. [MR1108330](#)
- GOWER, J. C. (1975). Generalized Procrustes analysis. *Psychometrika* **40** 33–51. [MR0405725](#)
- HUCKEMANN, S., HOTZ, T. and MUNK, A. (2010). Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statist. Sinica* **20** 1–100.
- HUCKEMANN, S. and ZIEZOLD, H. (2006). Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Adv. in Appl. Probab.* **38** 299–319. [MR2264946](#)
- JOLLIFFE, I. T. (2002). *Principal Component Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR2036084](#)

Key words and phrases. Dimension reduction, manifold, principal components analysis, principal nested spheres, Riemannian, shape.

- JUNG, S., DRYDEN, I. L. and MARRON, J. S. (2012). Analysis of principal nested spheres. *Biometrika* **99** 551–568. [MR2966769](#)
- KENDALL, D. G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *Bull. Lond. Math. Soc.* **16** 81–121. [MR0737237](#)
- KENOBI, K., DRYDEN, I. L. and LE, H. (2010). Shape curves and geodesic modelling. *Biometrika* **97** 567–584. [MR2672484](#)
- KENT, J. T. (1994). The complex Bingham distribution and shape analysis. *J. Roy. Statist. Soc. Ser. B* **56** 285–299. [MR1281934](#)
- KENT, J. T. and MARDIA, K. V. (2001). Shape, Procrustes tangent projections and bilateral symmetry. *Biometrika* **88** 469–485. [MR1844846](#)
- LE, H. L. (1991). On geodesics in Euclidean shape spaces. *J. Lond. Math. Soc.* (2) **44** 360–372. [MR1136446](#)
- LE, H. L. and KENDALL, D. G. (1993). The Riemannian structure of Euclidean shape spaces: A novel environment for statistics. *Ann. Statist.* **21** 1225–1271. [MR1241266](#)
- MARGULIS, C. J., STERN, H. A. and BERNE, B. J. (2002). Helix unfolding and intramolecular hydrogen bond dynamics in small α -helices in explicit solvent. *J. Phys. Chem. B* **106** 10748–10752.
- MARRON, J. S. and ALONSO, A. M. (2014). Overview of object oriented data analysis. *Biom. J.* **56** 732–753. [MR3258083](#)
- PANARETOS, V. M., PHAM, T. and YAO, Z. (2014). Principal flows. *J. Amer. Statist. Assoc.* **109** 424–436. [MR3180574](#)
- PENNEC, X. (2018). Barycentric subspace analysis on manifolds. *Ann. Statist.* **46** 2711–2746. [MR3851753](#)
- SALOMON-FERRER, R., CASE, D. A. and WALKER, R. C. (2013). An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **3** 198–210.
- WANG, H. and MARRON, J. S. (2007). Object oriented data analysis: Sets of trees. *Ann. Statist.* **35** 1849–1873. [MR2363955](#)
- WARD, J. H. JR. (1963). Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.* **58** 236–244. [MR0148188](#)

SPATIAL MODELING OF TRENDS IN CRIME OVER TIME IN PHILADELPHIA

BY CECILIA BALOCCHI AND SHANE T. JENSEN

University of Pennsylvania

Understanding the relationship between change in crime over time and the geography of urban areas is an important problem for urban planning. Accurate estimation of changing crime rates throughout a city would aid law enforcement as well as enable studies of the association between crime and the built environment. Bayesian modeling is a promising direction since areal data require principled sharing of information to address spatial autocorrelation between proximal neighborhoods. We develop several Bayesian approaches to spatial sharing of information between neighborhoods while modeling trends in crime counts over time. We apply our methodology to estimate changes in crime throughout Philadelphia over the 2006–15 period while also incorporating spatially-varying economic and demographic predictors. We find that the local shrinkage imposed by a conditional autoregressive model has substantial benefits in terms of out-of-sample predictive accuracy of crime. We also explore the possibility of spatial discontinuities between neighborhoods that could represent natural barriers or aspects of the built environment.

REFERENCES

- ALDOR-NOIMAN, S., BROWN, L. D., FOX, E. B. and STINE, R. A. (2016). Spatio-temporal low count processes with application to violent crime events. *Statist. Sinica* **26** 1587–1610. [MR3586230](#)
- ANDERSON, C. and RYAN, L. M. (2017). A comparison of spatio-temporal disease mapping approaches including an application to ischaemic heart disease in New South Wales, Australia. *International Journal of Environmental Research and Public Health* **14** 146.
- ANSELIN, L. (1998). *Spatial Econometrics: Methods and Models*. Springer, Dordrecht.
- BALOCCHI, C. and JENSEN, S. T. (2019). Supplement to “Spatial modeling of trends in crime over time in Philadelphia.” DOI:[10.1214/19-AOAS1280SUPP](#).
- BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. *Monographs on Statistics and Applied Probability* **135**. CRC Press, Boca Raton, FL. [MR3362184](#)
- BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. [MR0373208](#)
- BESAG, J., YORK, J. and MOLLIE, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43** 1–59. [MR1105822](#)
- BOOTS, B. (2001). Using local statistics for boundary characterization. *GeoJournal* **53** 339–345.
- BRANTINGHAM, P. J. and BRANTINGHAM, P. L. (1984). *Patterns in Crime*. Macmillan, New York.
- BROOK, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika* **51** 481–483. [MR0205315](#)

Key words and phrases. Urbanism, crime, spatial, time trends.

- BURBIDGE, J. B., MAGEE, L. and ROBB, A. L. (1988). Alternative transformations to handle extreme values of the dependent variable. *J. Amer. Statist. Assoc.* **83** 123–127. [MR0941005](#)
- CRESSIE, N. (1990). The origins of kriging. *Math. Geol.* **22** 239–252. [MR1047810](#)
- FBI. Uniform crime reporting program, definitions. Available at <https://ucr.fbi.gov/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/offense-definitions>. Accessed: 2016-09-15.
- FLAXMAN, S. R. (2014). A general approach to prediction and forecasting crime rates with Gaussian processes. Heinz College Technical Report, Carnegie Mellon Univ., Pittsburgh, PA.
- GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.
- HERBERT, D. T. (1982). *The Geography of Urban Crime*. Longman, London.
- HUMPHREY, C., JENSEN, S. T., SMALL, D. S. and THURSTON, R. (2017). Analysis of Urban Vibrancy and Safety in Philadelphia. Working paper. Available at [arXiv:1702.07909](https://arxiv.org/abs/1702.07909).
- HVISTENDAHL, M. (2016). Crime forecasters. *Science* **353** 1484–1487.
- JOHNSON, S. D., LAB, S. P. and BOWERS, K. J. (2008). Stable and fluid hotspots of crime: Differentiation and identification. *Built Environment* **34** 32–45.
- KAHLE, D. and WICKHAM, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal* **5**.
- LAW, J., QUICK, M. and CHAN, P. (2014). Bayesian spatio-temporal modeling for analysing local patterns of crime over time at the small-area level. *J. Quant. Criminol.* **30** 57–78.
- LEE, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spat Spatiotemporal Epidemiol* **2** 79–89.
- LEE, D. and MITCHELL, R. (2012). Boundary detection in disease mapping studies. *Biostatistics* **13** 415–426.
- LEE, D. and MITCHELL, R. (2013). Locally adaptive spatial smoothing using conditional autoregressive models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 593–608. [MR3083913](#)
- LEROUX, B. G., LEI, X. and BRESLOW, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (Minneapolis, MN, 1997). *IMA Vol. Math. Appl.* **116** 179–191. Springer, New York. [MR1731684](#)
- LI, P., BANERJEE, S. and MCBEAN, A. M. (2011). Mining boundary effects in areally referenced spatial data using the Bayesian information criterion. *Geoinformatica* **15** 435–454.
- LI, P., BANERJEE, S., MCBEAN, A. M. and CARLIN, B. P. (2012). Bayesian areal wombling using false discovery rates. *Stat. Interface* **5** 149–158. [MR2928066](#)
- LI, G., HAINING, R., RICHARDSON, S. and BEST, N. (2014). Space-time variability in burglary risk: A Bayesian spatio-temporal modelling approach. *Spat. Stat.* **9** 180–191. [MR3326838](#)
- LU, H. and CARLIN, B. P. (2005). Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis* **37** 265–285.
- 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](#)
- MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. [MR2816705](#)
- MØLLER, J. and WAAGEPETERSEN, R. P. (2007). Modern statistics for spatial point processes. *Scand. J. Stat.* **34** 643–684. [MR2392447](#)
- MORAN, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika* **37** 17–23. [MR0035933](#)
- PACE, R. K. and LESAGE, J. (2010). Spatial econometrics. In *Handbook of Spatial Statistics*. Chapman & Hall/CRC Handb. Mod. Stat. Methods 245–260. CRC Press, Boca Raton, FL. [MR2730939](#)
- STEIN, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer, New York. [MR1697409](#)
- TADDY, M. A. (2010). Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime. *J. Amer. Statist. Assoc.* **105** 1403–1417. [MR2796559](#)

WHITTLE, P. (1954). On stationary processes in the plane. *Biometrika* **41** 434–449. [MR0067450](#)
WOMBLE, W. H. (1951). Differential systematics. *Science* **114** 315–322.

FITTING A DEEPLY NESTED HIERARCHICAL MODEL TO A LARGE BOOK REVIEW DATASET USING A MOMENT-BASED ESTIMATOR

BY NINGSHAN ZHANG, KYLE SCHMAUS AND PATRICK O. PERRY

New York University, Stitch Fix and Oscar Health

We consider a particular instance of a common problem in recommender systems, using a database of book reviews to inform user-targeted recommendations. In our dataset, books are categorized into genres and subgenres. To exploit this nested taxonomy, we use a hierarchical model that enables information pooling across across similar items at many levels within the genre hierarchy. The main challenge in deploying this model is computational. The data sizes are large and fitting the model at scale using off-the-shelf maximum likelihood procedures is prohibitive. To get around this computational bottleneck, we extend a moment-based fitting procedure proposed for fitting single-level hierarchical models to the general case of arbitrarily deep hierarchies. This extension is an order of magnitude faster than standard maximum likelihood procedures. The fitting method can be deployed beyond recommender systems to general contexts with deeply nested hierarchical generalized linear mixed models.

REFERENCES

- ADOMAVICIUS, G. and TUZHILIN, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **17** 734–749.
- AGARWAL, D. (2008). Statistical challenges in Internet advertising. In *Statistical Methods in e-Commerce Research. Statist. Practice* 3–17. Wiley, Hoboken, NJ. [MR2503567](#)
- 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. WSDM '10* 91–100. ACM, New York.
- ANSARI, A., ESSEGAIER, S. and KOHLI, R. (2000). Internet recommendation systems. *J. Mark. Res.* **37** 363–375.
- BATES, D., MAECHLER, M., BOLKER, B. and WALKER, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7.
- COCHRAN, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Suppl. J. R. Stat. Soc.* **4** 102–118.
- CONDLIFF, M. K., LEWIS, D. D. and MADIGAN, D. (1999). Bayesian mixed-effects models for recommender systems. In *ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation*.
- DROR, G., KOENIGSTEIN, N. and KOREN, Y. (2011). Yahoo! Music recommendations: Modeling music ratings with temporal dynamics and item taxonomy. In *Proceedings of the Fifth ACM Conference on Recommender Systems* 165–172. ACM, New York.

Key words and phrases. Generalized linear mixed model, hierarchical model, method of moments, recommender system.

- FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80** 27–38. [MR1225212](#)
- GAO, K. and OWEN, A. B. (2016). Estimation and inference for very large linear mixed effects models. ArXiv E-prints.
- GAO, K. and OWEN, A. (2017). Efficient moment calculations for variance components in large unbalanced crossed random effects models. *Electron. J. Stat.* **11** 1235–1296. [MR3635913](#)
- GEBREGZIABHER, M., EGEDE, L., GILBERT, G. E., HUNT, K., NIETERT, P. J. and MAULDIN, P. (2012). Fitting parametric random effects models in very large data sets with application to VHA national data. *BMC Med. Res. Methodol.* **12** 1–14.
- HUANG, Z. and GELMAN, A. (2005). Sampling for Bayesian computation with large datasets. Unpublished.
- KOHAVI, R., DENG, A., LONGBOTHAM, R. and XU, Y. (2014). Seven rules of thumb for web site experimenters. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1857–1866. ACM, New York.
- KOREN, Y., BELL, R. and VOLINKSY, C. (2009). Matrix factorization techniques for recommender systems. *Computer* **42** 30–37.
- KRAMER, A. D., GUILLORY, J. E. and HANCOCK, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. USA* **111** 8788–8790.
- MCAULEY, J., PANDEY, R. and LESKOVEC, J. (2015). Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794.
- NAIK, P., WEDEL, M., BACON, L., BODAPATI, A., BRADLOW, E., KAMAKURA, W., KREULEN, J., LENK, P., MADIGAN, D. M. et al. (2008). Challenges and opportunities in high-dimensional choice data analyses. *Mark. Lett.* **19** 201–213.
- PERRY, P. O. (2017). Fast moment-based estimation for hierarchical models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 267–291. [MR3597973](#)
- SCOTT, S. L., BLOCKER, A. W., BONASSI, F. V., CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2013). Bayes and big data: The consensus Monte Carlo algorithm. In *Bayes* 250.
- TAN, Z., ROCHE, K., ZHOU, X. and MUKHERJEE, S. (2018). Scalable algorithms for learning high-dimensional linear mixed models. Available at [arXiv:1803.04431](#).
- WENG, L. T., XU, Y., LI, Y. and NAYAK, R. (2008). Exploiting item taxonomy for solving cold-start problem in recommendation making. In *2008 20th IEEE International Conference on Tools with Artificial Intelligence* **2** 113–120.
- ZHANG, Y., CAO, B. and YEUNG, D.-Y. (2010). Multi-domain collaborative filtering. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence. UAI'10* 725–732. AUAI Press, Arlington, VA.
- ZHANG, X., ZHOU, Y., MA, Y., CHEN, B.-C., ZHANG, L. and AGARWAL, D. (2016). GLMix: Generalized linear mixed models for large-scale response prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16* 363–372. ACM, New York.
- ZHOU, X., CARBONETTO, P. and STEPHENS, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* **9** e1003264.
- ZIEGLER, C.-N., MCNEE, S. M., KONSTAN, J. A. and LAUSEN, G. (2005). Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web* 22–32. ACM, New York.

OBJECTIVE BAYES MODEL SELECTION OF GAUSSIAN INTERVENTIONAL ESSENTIAL GRAPHS FOR THE IDENTIFICATION OF SIGNALING PATHWAYS¹

BY FEDERICO CASTELLETTI AND GUIDO CONSONNI

Università Cattolica del Sacro Cuore

A signalling pathway is a sequence of chemical reactions initiated by a stimulus which in turn affects a receptor, and then through some intermediate steps cascades down to the final cell response. Based on the technique of flow cytometry, samples of cell-by-cell measurements are collected under each experimental condition, resulting in a collection of interventional data (assuming no latent variables are involved). Usually several external interventions are applied at different points of the pathway, the ultimate aim being the structural recovery of the underlying signalling network which we model as a causal Directed Acyclic Graph (DAG) using intervention calculus. The advantage of using interventional data, rather than purely observational one, is that identifiability of the true data generating DAG is enhanced. More technically a Markov equivalence class of DAGs, whose members are statistically indistinguishable based on observational data alone, can be further decomposed, using additional interventional data, into smaller distinct Interventional Markov equivalence classes. We present a Bayesian methodology for structural learning of Interventional Markov equivalence classes based on observational and interventional samples of multivariate Gaussian observations. Our approach is objective, meaning that it is based on default parameter priors requiring no personal elicitation; some flexibility is however allowed through a tuning parameter which regulates sparsity in the prior on model space. Based on an analytical expression for the marginal likelihood of a given Interventional Essential Graph, and a suitable MCMC scheme, our analysis produces an approximate posterior distribution on the space of Interventional Markov equivalence classes, which can be used to provide uncertainty quantification for features of substantive scientific interest, such as the posterior probability of inclusion of selected edges, or paths.

REFERENCES

- ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997a). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25** 505–541. [MR1439312](#)
- ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997b). On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scand. J. Stat.* **24** 81–102. [MR1436624](#)
- ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Stat.* **28** 33–85. [MR1844349](#)
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. [MR1325392](#)

Key words and phrases. Graphical model, causal inference, Markov equivalence class.

- CAO, X., KHARE, K. and GHOSH, M. (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *Ann. Statist.* **47** 319–348. [MR3909935](#)
- CASTELLETTI, F. and CONSONNI, G. (2019). Supplement to “Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways.” DOI:[10.1214/19-AOAS1275SUPP](#).
- CASTELLETTI, F., CONSONNI, G., DELLA VEDOVA, M. L. and PELUSO, S. (2018). Learning Markov equivalence classes of directed acyclic graphs: An objective Bayes approach. *Bayesian Anal.* **13** 1231–1256. [MR3855370](#)
- CHICKERING, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* **2** 445–498. [MR1929415](#)
- CONSONNI, G., LA ROCCA, L. and PELUSO, S. (2017). Objective Bayes covariate-adjusted sparse graphical model selection. *Scand. J. Stat.* **44** 741–764. [MR3687971](#)
- COWELL, R. G., DAWID, A. P., LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1999). *Probabilistic Networks and Expert Systems. Statistics for Engineering and Information Science*. Springer, New York. [MR1697175](#)
- FOYGEL, R. and DRTON, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. *Adv. Neural Inf. Process. Syst.* **23** 2020–2028.
- FRIEDMAN, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* **303** 799–805.
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.
- GEIGER, D. and HECKERMAN, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30** 1412–1440. [MR1936324](#)
- GEISSER, S. and CORNFELD, J. (1963). Posterior distributions for multivariate normal parameters. *J. Roy. Statist. Soc. Ser. B* **25** 368–376. [MR0171354](#)
- GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. [MR1422404](#)
- GILLISPIE, S. B. and PERLMAN, M. D. (2002). The size distribution for Markov equivalence classes of acyclic digraph models. *Artificial Intelligence* **141** 137–155. [MR1935281](#)
- HAUSER, A. and BÜHLMANN, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.* **13** 2409–2464. [MR2973606](#)
- HAUSER, A. and BÜHLMANN, P. (2015). Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 291–318. [MR3299409](#)
- HE, Y.-B. and GENG, Z. (2008). Active learning of causal networks with intervention experiments and optimal designs. *J. Mach. Learn. Res.* **9** 2523–2547. [MR2460892](#)
- HE, Y., JIA, J. and YU, B. (2013). Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. *Ann. Statist.* **41** 1742–1779. [MR3127848](#)
- HOIJTINK, H. (2013). Objective Bayes factors for inequality constrained hypotheses. *Int. Stat. Rev.* **81** 207–229. [MR3100657](#)
- LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. [MR1419991](#)
- LUO, R. and ZHAO, H. (2011). Bayesian hierarchical modeling for signaling pathway inference from single cell interventional data. *Ann. Appl. Stat.* **5** 725–745. [MR2840173](#)
- MAATHUIS, M. H., KALISCH, M. and BÜHLMANN, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.* **37** 3133–3164. [MR2549555](#)
- NESS, R. O., SACHS, K., MALLICK, P. and VITEK, O. (2017). A Bayesian active learning experimental design for inferring signaling networks. In *Research in Computational Molecular Biology. Lecture Notes in Computer Science* **10229** 134–156. Springer, Cham. [MR3662679](#)

- PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710. [MR1380809](#)
- PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press, Cambridge. [MR1744773](#)
- PETERS, J. and BÜHLMANN, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* **101** 219–228. [MR3180667](#)
- PETERSON, C., STINGO, F. C. and VANNUCCI, M. (2015). Bayesian inference of multiple Gaussian graphical models. *J. Amer. Statist. Assoc.* **110** 159–174. [MR3338494](#)
- RICHARDSON, T. and SPIRITES, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030. [MR1926166](#)
- SACHS, K., PEREZ, O., PE’ER, D., LAUFFENBURGER, D. and NOLAN, G. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308** 523–529.
- SHOJAIE, A. and MICHAILIDIS, G. (2009). Analysis of gene sets based on the underlying regulatory network. *J. Comput. Biol.* **16** 407–426. [MR2487566](#)
- SPIRITES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR1815675](#)
- TAN, L. S. L., JASRA, A., DE IORIO, M. and EBBELS, T. M. D. (2017). Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. *Ann. Appl. Stat.* **11** 2222–2251. [MR3743295](#)
- VERMA, T. and PEARL, J. (1991). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, UAI 90* 255–270. Elsevier, New York.

A LATENT DISCRETE MARKOV RANDOM FIELD APPROACH TO IDENTIFYING AND CLASSIFYING HISTORICAL FOREST COMMUNITIES BASED ON SPATIAL MULTIVARIATE TREE SPECIES COUNTS¹

BY STEPHEN BERG, JUN ZHU, MURRAY K. CLAYTON, MONIKA E. SHEA
 AND DAVID J. MLADENOFF

University of Wisconsin-Madison

The Wisconsin Public Land Survey database describes historical forest composition at high spatial resolution and is of interest in ecological studies of forest composition in Wisconsin just prior to significant Euro-American settlement. For such studies it is useful to identify recurring subpopulations of tree species known as communities, but standard clustering approaches for subpopulation identification do not account for dependence between spatially nearby observations. Here, we develop and fit a latent discrete Markov random field model for the purpose of identifying and classifying historical forest communities based on spatially referenced multivariate tree species counts across Wisconsin. We show empirically for the actual dataset and through simulation that our latent Markov random field modeling approach improves prediction and parameter estimation performance. For model fitting we introduce a new stochastic approximation algorithm which enables computationally efficient estimation and classification of large amounts of spatial multivariate count data.

REFERENCES

- BARNES, B., ZAK, D., DENTON, S. and SPURR, S. (2010). *Forest Ecology*, 4th ed. Wiley, New York.
- BENVENISTE, A., PRIOURET, P. and MÉTIVIER, M. (1990). *Adaptive Algorithms and Stochastic Approximations. Applications of Mathematics (New York)* **22**. Springer, Berlin. [MR1082341](#)
- BERG, S., ZHU, J., CLAYTON, M. K., SHEA, M. E and MLADENOFF, D. J (2019). Supplement to “A latent discrete Markov random field approach to identifying and classifying historical forest communities based on spatial multivariate tree species counts.” DOI:[10.1214/19-AOAS1259SUPP](#).
- BURNS, R. M. and HONKALA, B. H. (1990). *Silvics of North America* **2**. U.S. Department of Agriculture, Washington, DC.
- CHEN, J. (2017). Consistency of the MLE under mixture models. *Statist. Sci.* **32** 47–63. [MR3634306](#)
- COMETS, F. and GIDAS, B. (1992). Parameter estimation for Gibbs distributions from partially observed data. *Ann. Appl. Probab.* **2** 142–170. [MR1143397](#)
- CURTIS, J. T. (1959). *The Vegetation of Wisconsin: An Ordination of Wisconsin Plant Communities*. Univ. Wisconsin Press, Madison, WI.
- DAVIS, M. B., SCHWARTZ, M. W. and WOODS, K. (1991). Detecting a species limit from pollen in sediments. *J. Biogeogr.* **18** 653–668.

Key words and phrases. Markov random field, historical ecology, stochastic approximation, maximum likelihood, mixture models.

- 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](#)
- EGAN, D. (2005). *The Historical Ecology Handbook: A Restorationist's Guide to Reference Ecosystems*. Island Press.
- FAHEY, R. T., LORIMER, C. G. and MLADENOFF, D. J. (2012). Habitat heterogeneity and life-history traits influence presettlement distributions of early-successional tree species in a late-successional, hemlock-hardwood landscape. *Landsc. Ecol.* **27** 999–1013.
- FINLEY, R. W. (1976). The original vegetation cover of Wisconsin. *Wisconsin Department of Natural Resources*.
- FORBES, F. and FORT, G. (2007). Combining Monte Carlo and mean-field-like methods for inference in hidden Markov random fields. *IEEE Trans. Image Process.* **16** 824–837. [MR2460196](#)
- FORBES, F., CHARRAS-GARRIDO, M., AZIZI, L., DOYLE, S. and ABRIAL, D. (2013). Spatial risk mapping for rare disease with hidden Markov fields and variational EM. *Ann. Appl. Stat.* **7** 1192–1216. [MR3113506](#)
- FORT, G. and MOULINES, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Statist.* **31** 1220–1259. [MR2001649](#)
- GAETAN, C. and GUYON, X. (2010). *Spatial Statistics and Modeling. Springer Series in Statistics*. Springer, New York. [MR2569034](#)
- GANGNON, R. E. and CLAYTON, M. K. (2003). A hierarchical model for spatially clustered disease rates. *Stat. Med.* **22** 3213–3228.
- GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185. [MR1647507](#)
- GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-6** 721–741.
- HONG, C., NING, Y., WANG, S., WU, H., CARROLL, R. J. and CHEN, Y. (2017). PLEMT: A novel pseudolikelihood-based EM test for homogeneity in generalized exponential tilt mixture models. *J. Amer. Statist. Assoc.* **112** 1393–1404. [MR3750863](#)
- KNORR-HELD, L. and RASSER, G. (2004). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56** 13–21.
- KUSHNER, H. J. and YIN, G. G. (1997). *Stochastic Approximation Algorithms and Applications. Applications of Mathematics (New York)* **35**. Springer, New York. [MR1453116](#)
- LAWSON, A. B. (2010). Hotspot detection and clustering: Ways and means. *Environ. Ecol. Stat.* **17** 231–245. [MR2725781](#)
- LIU, F., MLADENOFF, D. J., KEULER, N. S. and MOORE, L. S. (2011). BROADSCALE variability in tree data of the historical Public Land Survey and its consequences for ecological studies. *Ecol. Monogr.* **81** 259–275.
- MLADENOFF, D. J., SICKLEY, T. A., SCHULTE, L. A., RHEMTULLA, J. M. and BOLLIGER, J. (2009). Wisconsin's Land Cover in the 1800s. *Wisconsin Department of Natural Resources*.
- MLADENOFF, D. J., WHITE, M. A., PASTOR, J. and CROW, T. R. (1993). Comparing spatial pattern in unaltered old-growth and disturbed forest landscapes. *Ecol. Appl.* **3** 294–306.
- NEAL, R. M. (1993). Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report.
- PACIOREK, C. J., GORING, S. J., THURMAN, A. L., COGBILL, C. V., WILLIAMS, J. W., MLADENOFF, D. J., PETERS, J. A., ZHU, J. and MCLACHLAN, J. S. (2016). Statistically-estimated tree composition for the northeastern United States at Euro-American settlement. *PLoS ONE* **11** 1–20.
- PETERSON, D. W. and REICH, P. B. (2001). Prescribed fire in oak savanna: Fire frequency effects on stand structure and dynamics. *Ecol. Appl.* **11** 914–927.
- RADELOFF, V. C., MLADENOFF, D. J., HE, H. S. and BOYCE, M. S. (1999). Forest landscape change in the northwestern Wisconsin Pine Barrens from pre-European settlement to the present. *Can. J. For. Res.* **29** 1649–1659.

- ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. [MR0042668](#)
- SCHULTE, L. and MLADENOFF, D. (2001). The original US public land survey records: Their use and limitations in reconstructing pre-European settlement vegetation. *J. For.* **99** 5–10.
- SCHULTE, L. A., MLADENOFF, D. J. and NORDHEIM, E. V. (2002). Quantitative classification of a historic northern Wisconsin (USA) landscape: Mapping forests at regional scales. *Can. J. For. Res.* **32** 1616–1638.
- SHAO, J. (2003). *Mathematical Statistics*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. [MR2002723](#)
- SHEA, M. E., SCHULTE, L. A. and PALIK, B. J. (2014). Reconstructing vegetation past: Pre-European vegetation for the midwest Driftless area, USA. *Ecol. Restor.* **32** 417–433.
- STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ_1 -penalization for mixture regression models. *TEST* **19** 209–256. [MR2677722](#)
- STAMBAUGH, M. C. and GUYETTE, R. P. (2008). Predicting spatio-temporal variability in fire return intervals using a topographic roughness index. *For. Ecol. Manag.* **254** 463–473.
- WALLER, L. A. (2009). Detection of clustering in spatial data. In *The SAGE Handbook of Spatial Analysis* (J. Fagerberg, D. C. Mowery and R. R. Nelson, eds.) 299–321 10. Sage, London.
- WU, F. Y. (1982). The Potts model. *Rev. Modern Phys.* **54** 235–268. [MR0641370](#)
- WU, C.-F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103. [MR0684867](#)
- YOUNES, L. (1989). Parametric inference for imperfectly observed Gibbsian fields. *Probab. Theory Related Fields* **82** 625–645. [MR1002904](#)

A NONPARAMETRIC SPATIAL TEST TO IDENTIFY FACTORS THAT SHAPE A MICROBIOME

BY SUSHEELA P. SINGH*, ANA-MARIA STAICU*, ROBERT R. DUNN*,
 NOAH FIERER[†] AND BRIAN J. REICH*

*North Carolina State University** and *University of Colorado[†]*

The advent of high-throughput sequencing technologies has made data from DNA material readily available, leading to a surge of microbiome-related research establishing links between markers of microbiome health and specific outcomes. However, to harness the power of microbial communities we must understand not only how they affect us, but also how they can be influenced to improve outcomes. This area has been dominated by methods that reduce community composition to summary metrics, which can fail to fully exploit the complexity of community data. Recently, methods have been developed to model the abundance of taxa in a community, but they can be computationally intensive and do not account for spatial effects underlying microbial settlement. These spatial effects are particularly relevant in the microbiome setting because we expect communities that are close together to be more similar than those that are far apart. In this paper, we propose a flexible Bayesian spike-and-slab variable selection model for presence-absence indicators that accounts for spatial dependence and cross-dependence between taxa while reducing dimensionality in both directions. We show by simulation that in the presence of spatial dependence, popular distance-based hypothesis testing methods fail to preserve their advertised size, and the proposed method improves variable selection. Finally, we present an application of our method to an indoor fungal community found within homes across the contiguous United States.

REFERENCES

- ANDERSON, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26** 32–46.
- BANERJEE, S. (2005). On geodetic distance computations in spatial modeling. *Biometrics* **61** 617–625. [MR2140936](#)
- BARBERÁN, A., DUNN, R. R., REICH, B. J., PACIFICI, K., LABER, E. B., MENNINGER, H. L., MORTON, J. M., HENLEY, J. B., LEFF, J. W. et al. (2015). The ecology of microscopic life in household dust. *Proc. R. Soc. Lond., B Biol. Sci.* **282** 212–220.
- BRAY, J. R. and CURTIS, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* **27** 325–349.
- CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. [MR3059077](#)
- CHEN, J. and LI, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* **7** 418–442. [MR3086425](#)

Key words and phrases. Bayesian nonparametrics, Dirichlet process, high dimensional data, spatial modeling, spike-and-slab prior, variable selection.

- CLARK, J. S., NEMERGUT, D., SEYEDNASROLLAH, B., TURNER, P. J. and ZHANG, S. (2017). Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecol. Monogr.* **87** 34–56.
- CLARKE, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Aust. J. Ecol.* **18** 117–143.
- CRAVEN, P. and WAHBA, G. (1978). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–403.
- DANNEMILLER, K. C., GENT, J. F., LEADERER, B. P. and PECCIA, J. (2016). Influence of housing characteristics on bacterial and fungal communities in homes of asthmatic children. *Indoor Air* **26** 179–192.
- DUNN, R. R., FIERER, N., HENLEY, J. B., LEFF, J. W. and MENNINGER, H. L. (2013). Home life: Factors structuring the bacterial diversity found within and between homes. *PLoS ONE* **8** e64133.
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. [MR0350949](#)
- FLORES, G. E., HENLEY, J. B. and FIERER, N. (2012). A direct PCR approach to accelerate analyses of human-associated microbial communities. *PLoS ONE* **7** e44563.
- FRY, J. A., XIAN, G., JIN, S., DEWITZ, J. A., HOMER, C. G., LIMIN, Y., BARNES, C. A., HEROLD, N. D. and WICKHAM, J. D. (2011). Completion of the 2006 national land cover database for the conterminous United States. *Photogramm. Eng. Remote Sens.* **77** 858–864.
- GELFAND, A. E., KOTTAS, A. and MACEACHERN, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100** 1021–1035. [MR2201028](#)
- GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.
- GRANTHAM, N. S., REICH, B. J., PACIFICI, K., LABER, E. B., MENNINGER, H. L., HENLEY, J. B., BARBERÁN, A., LEFF, J. W., FIERER, N. et al. (2015). Fungi identify the geographic origin of dust samples. *PLoS ONE* **10** e0122605.
- GRANTHAM, N. S., REICH, B. J., BORER, E. T. and GROSS, K. (2017). MIMIX: A Bayesian mixed-effects model for microbiome data from designed experiments. Manuscript in review.
- HALL, P., MÜLLER, H.-G. and YAO, F. (2008). Modelling sparse generalized longitudinal observations with latent Gaussian processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 703–723. [MR2523900](#)
- HAMADA, N. and FUJITA, T. (2002). Effect of air-conditioner on fungal contamination. *Atmos. Environ.* **36** 5443–5448.
- HARRIS, I., JONES, P. D., OSBORN, T. J. and LISTER, D. H. (2014). Updated high-resolution grids of monthly climatic observations—the CRU TS3.10 dataset. *Int. J. Climatol.* **34** 623–642.
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR2722294](#)
- HEATHER, J. M. and CHAIN, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics* **107** 1–8.
- HUMAN MICROBIOME PROJECT CONSORTIUM (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486** 207–214.
- KETTLESON, E. M., ADHIKARI, A., VESPER, S., COOMBS, K., INDUGULA, R. and REPONEN, T. (2015). Key determinants of the fungal and bacterial microbiomes in homes. *Environ. Res.* **138** 130–135.
- KUO, L. and MALLICK, B. (1998). Variable selection for regression models. *Sankhya B* **60** 65–81. [MR1717076](#)
- LEE, S., HUANG, J. Z. and HU, J. (2010). Sparse logistic principal components analysis for binary data. *Ann. Appl. Stat.* **4** 1579–1601. [MR2758342](#)
- LORENZ, E. N. (1956). Empirical orthogonal functions and statistical weather prediction.

- MCARDLE, B. H. and ANDERSON, M. J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology* **82** 290–297.
- MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. With comments by James Berger and C. L. Mallows and with a reply by the authors. [MR0997578](#)
- NELSEN, R. B. (1999). *An Introduction to Copulas. Lecture Notes in Statistics* **139**. Springer, New York. [MR1653203](#)
- OVASKAINEN, O., HOTTOLA, J. and SIITONEN, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology* **91** 2514–2521.
- OVASKAINEN, O., ROY, D. B., FOX, R. and ANDERSON, B. J. (2016). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods Ecol. Evol.* **7** 428–436.
- OVASKAINEN, O., TIKHONOV, G., NORBERG, A., GUILLAUME BLANCHET, F., DUAN, L., DUNSON, D., ROSLIN, T. and ABREGO, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* **20** 561–576.
- PETRONE, S., GUINDANI, M. and GELFAND, A. E. (2009). Hybrid Dirichlet mixture models for functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 755–782. [MR2750094](#)
- QIN, J., LI, Y., CAI, Z., LI, S., ZHU, J., ZHANG, F., LIANG, S., ZHANG, W., GUAN, Y. et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490** 55–60.
- RAVEL, J., GAJER, P., ABDO, Z., SCHNEIDER, G. M., KOENIG, S. S. K., MCCULLE, S. L., KARLEBACH, S., GORLE, R., RUSSELL, J. et al. (2011). Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. USA* **108** 4680–4687.
- REICH, B. J. and FUENTES, M. (2007). A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Ann. Appl. Stat.* **1** 249–264. [MR2393850](#)
- REUTER, J. A., SPACEK, D. V. and SNYDER, M. P. (2015). High-throughput sequencing technologies. *Molecular Cell* **58** 586–597.
- ROČKOVÁ, V. and GEORGE, E. I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.* **113** 431–444. [MR3803476](#)
- RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2010). Latent stick-breaking processes. *J. Amer. Statist. Assoc.* **105** 647–659. [MR2724849](#)
- ROUND, J. L. and MAZMANIAN, S. K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nat. Rev., Immunol.* **9** 313–323.
- SERBAN, N., STAIU, A.-M. and CARROLL, R. J. (2013). Multilevel cross-dependent binary longitudinal data. *Biometrics* **69** 903–913. [MR3146786](#)
- SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. [MR1309433](#)
- SHIROTA, S., GELFAND, A. E. and BANERJEE, S. (2017). Spatial joint species distribution modeling using Dirichlet processes.
- SINGH, S. P., STAIU, A., DUNN, R. R., FIERER, N. and REICH, B. J. (2019). Supplement to “A nonparametric spatial test to identify factors that shape a microbiome.” DOI:[10.1214/19-AOAS1262SUPP](#).
- STEIN, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging. Springer Series in Statistics*. Springer, New York. [MR1697409](#)
- THORSON, J. T., SCHEUERELL, M. D., SHELTON, A. O., SEE, K. E., SKAUG, H. J. and KRISTENSEN, K. (2015). Spatial factor analysis: A new tool for estimating joint species distributions and correlations in species range. *Methods Ecol. Evol.* **6** 627–637.
- TURNBAUGH, P. J., HAMADY, M., YATSUNENKO, T., CANTAREL, B. L., DUNCAN, A., LEY, R. E., SOGIN, M. L., JONES, W. J., ROE, B. A. et al. (2009). A core gut microbiome in obese and lean twins. *Nature* **457** 480–484.

- WADSWORTH, W. D., ARGIENTO, R., GUINDANI, M., GALLOWAY-PENA, J., SHELBURNE, S. A. and VANNUCCI, M. (2017). An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinform.* **18** 94.
- WANG, T. and ZHAO, H. (2017). A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics* **73** 792–801. [MR3713113](#)
- WARTON, D. I. (2011). Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics* **67** 116–123. [MR2898823](#)
- WARTON, D. I., WRIGHT, S. T. and WANG, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. *Methods Ecol. Evol.* **3** 89–101.
- ZHAO, N., CHEN, J., CARROLL, I. M., RINGEL-KULKA, T., EPSTEIN, M. P., ZHOU, H., ZHOU, J. J., RINGEL, Y., LI, H. et al. (2015). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.* **96** 797–807.
- ZHOU, J., BHATTACHARYA, A., HERRING, A. H. and DUNSON, D. B. (2015). Bayesian factorizations of big sparse tensors. *J. Amer. Statist. Assoc.* **110** 1562–1576. [MR3449055](#)

PREDICTING PALEOCLIMATE FROM COMPOSITIONAL DATA USING MULTIVARIATE GAUSSIAN PROCESS INVERSE PREDICTION¹

BY JOHN R. TIPTON^{*}, MEVIN B. HOOTEN^{†,‡}, CONNOR NOLAN[§],
ROBERT K. BOOTH[¶] AND JASON MCLACHLAN^{||}

University of Arkansas^{}, Colorado State University[†], U. S. Geological Survey,
Colorado Cooperative Fish and Wildlife Research Unit[‡], University of Arizona[§],
Lehigh University[¶] and University of Notre Dame^{||}*

Multivariate compositional count data arise in many applications including ecology, microbiology, genetics and paleoclimate. A frequent question in the analysis of multivariate compositional count data is what underlying values of a covariate(s) give rise to the observed composition. Learning the relationship between covariates and the compositional count allows for inverse prediction of unobserved covariates given compositional count observations. Gaussian processes provide a flexible framework for modeling functional responses with respect to a covariate without assuming a functional form. Many scientific disciplines use Gaussian process approximations to improve prediction and make inference on latent processes and parameters. When prediction is desired on unobserved covariates given realizations of the response variable, this is called inverse prediction. Because inverse prediction is often mathematically and computationally challenging, predicting unobserved covariates often requires fitting models that are different from the hypothesized generative model. We present a novel computational framework that allows for efficient inverse prediction using a Gaussian process approximation to generative models. Our framework enables scientific learning about how the latent processes co-vary with respect to covariates while simultaneously providing predictions of missing covariates. The proposed framework is capable of efficiently exploring the high dimensional, multi-modal latent spaces that arise in the inverse problem. To demonstrate flexibility, we apply our method in a generalized linear model framework to predict latent climate states given multivariate count data. Based on cross-validation, our model has predictive skill competitive with current methods while simultaneously providing formal, statistical inference on the underlying community dynamics of the biological system previously not available.

REFERENCES

- AMESBURY, M., BARBER, K. and HUGHES, P. (2012). The relationship of fine-resolution, multi-proxy palaeoclimate records to meteorological data at Fågelmosse, Värmland, Sweden and the implications for the debate on climate drivers of the peat-based record. *Quat. Int.* **268** 77–86.

Key words and phrases. Bayesian hierarchical models, predictive validation, model comparison, ecological functional response model.

- ARMAGAN, A., DUNSON, D. B. and CLYDE, M. (2011). Generalized Beta mixtures of Gaussians. *Adv. Neural Inf. Process. Syst.* **24** 523–531.
- 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](#)
- BARNARD, J., MCCULLOCH, R. and MENG, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist. Sinica* **10** 1281–1311. [MR1804544](#)
- BHATTACHARYA, S. (2006). A Bayesian semiparametric model for organism based environmental reconstruction. *Environmetrics* **17** 763–776. [MR2252330](#)
- BIRKS, H. J. B. and SIMPSON, G. L. (2013). ‘Diatoms and pH reconstruction’ (1990) revisited. *J. Paleolimnol.* **49** 363–371.
- BIRKS, H. J. B., LINE, J. M., JUGGINS, S., STEVENSON, A. C. and TER BRAAK, C. J. F. (1990). Diatoms and pH reconstruction. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **327** 263–278.
- BOOTH, R. K. (2008). Testate amoebae as proxies for mean annual water-table depth in Sphagnum-dominated peatlands of North America. *J. Quat. Sci.* **23** 43–57.
- BOOTH, R. K., LAMENTOWICZ, M. and CHARMAN, D. J. (2010). Preparation and analysis of testate amoebae in peatland paleoenvironmental studies. *Mires Peat* **7** 1–7.
- BREWER, S., JACKSON, S. T. and WILLIAMS, J. W. (2012). Paleoecoinformatics: Applying geo-historical data to ecological questions. *Trends Ecol. Evol.* **27** 104–112.
- CHARMAN, D. J. (2007). Summer water deficit variability controls on peatland water-table changes: Implications for Holocene palaeoclimate reconstructions. *Holocene* **17** 217–227.
- 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](#)
- CSATÓ, L. (2002). Gaussian processes: Iterative sparse approximations. Ph.D. thesis, Aston Univ.
- 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](#)
- DAWSON, A., PACIOREK, C. J., MCLACHLAN, J. S., GORING, S., WILLIAMS, J. W. and JACKSON, S. T. (2016). Quantifying pollen-vegetation relationships to reconstruct ancient forests using 19th-century forest composition and pollen data. *Quat. Sci. Rev.* **137** 156–175.
- EDDELBUEITTEL, D. and SANDERSON, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Comput. Statist. Data Anal.* **71** 1054–1063. [MR3132026](#)
- FINLEY, A. O., SANG, H., BANERJEE, S. and GELFAND, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Comput. Statist. Data Anal.* **53** 2873–2884. [MR2667597](#)
- FUENTES, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *J. Amer. Statist. Assoc.* **102** 321–331. [MR2345545](#)
- GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. [MR1141740](#)
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. [MR2221284](#)
- GELMAN, A., RUBIN, D. B. et al. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
- GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. [MR2847988](#)
- 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](#)
- GRANTHAM, N. S., REICH, B. J., PACIFICI, K., LABER, E. B., MENNINGER, H. L., HENLEY, J. B., BARBERÁN, A., LEFF, J. W., FIERER, N. et al. (2015). Fungi identify the geographic origin of dust samples. *PLoS ONE* **10** e0122605.

- HASLETT, J., WHILEY, M., BHATTACHARYA, S., SALTER-TOWNSHEND, M., WILSON, S. P., ALLEN, J. R. M., HUNTLEY, B. and MITCHELL, F. J. G. (2006). Bayesian palaeoclimate reconstruction. *J. Roy. Statist. Soc. Ser. A* **169** 395–438. [MR2236914](#)
- HEFLEY, T. J., BROMS, K. M., BROST, B. M., BUDERMAN, F. E., KAY, S. L., SCHARF, H. R., TIPTON, J. R., WILLIAMS, P. J. and HOOTEN, M. B. (2017). The basis function approach for modeling autocorrelation in ecological data. *Ecology* **98** 632–646.
- HIGDON, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* 37–56. Springer, London. [MR2059819](#)
- JACKSON, S. T. and WILLIAMS, J. W. (2004). Modern analogs in Quaternary paleoecology: Here today, gone yesterday, gone tomorrow? *Annual Rev. Earth Planet. Sci.* **32** 495–537.
- JUGGINS, S. (2015). rioja: Analysis of quaternary science data. R package version 0.9-9.
- KAMMANN, E. E. and WAND, M. P. (2003). Geoadditive models. *J. Roy. Statist. Soc. Ser. C* **52** 1–18. [MR1963210](#)
- LAUBER, C. L., HAMADY, M., KNIGHT, R. and FIERER, N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* **75** 5111–5120.
- LEWANDOWSKI, D., KUROWICKA, D. and JOE, H. (2009). Generating random correlation matrices based on vines and extended onion method. *J. Multivariate Anal.* **100** 1989–2001. [MR2543081](#)
- LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. [MR2853727](#)
- MORALES-CASTILLA, I., MATIAS, M. G., GRAVEL, D. and ARAÚJO, M. B. (2015). Inferring biotic interactions from proxies. *Trends Ecol. Evol.* **30** 347–356.
- MURRAY, I., ADAMS, R. P. and MACKAY, D. J. C. (2010). Elliptical slice sampling. In *AISTATS* **13** 541–548.
- NOLAN, C., TIPTON, J., BOOTH, R. K., HOOTEN, M. B. and JACKSON, S. T. (2019). Comparing and improving methods for reconstructing peatland water-table depth from testate amoebae. *Holocene*.
- NYCHKA, D., BANDYOPADHYAY, S., HAMMERLING, D., LINDGREN, F. and SAIN, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *J. Comput. Graph. Statist.* **24** 579–599. [MR3357396](#)
- O'MALLEY, A. J. and ZASLAVSKY, A. M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *J. Amer. Statist. Assoc.* **103** 1405–1418. [MR2655721](#)
- OVERPECK, J., WEBB, T. and PRENTICE, I. (1985). Quantitative interpretation of fossil pollen spectra: Dissimilarity coefficients and the method of modern analogs. *Quat. Res.* **23** 87–108.
- OVERPECK, J. T., WEBB, R. S. and WEBB III, T. (1992). Mapping eastern North American vegetation change of the past 18 ka: No-analogs and the future. *Geology* **20** 1071–1074.
- PACIOREK, C. J. (2007). Computational techniques for spatial logistic regression with large data sets. *Comput. Statist. Data Anal.* **51** 3631–3653. [MR2364480](#)
- PACIOREK, C. J. and MCLACHLAN, J. S. (2009). Mapping ancient forests: Bayesian inference for spatio-temporal trends in forest composition using the fossil pollen proxy record. *J. Amer. Statist. Assoc.* **104** 608–622. [MR2751442](#)
- PARNELL, A. C., SWEENEY, J., DOAN, T. K., SALTER-TOWNSHEND, M., ALLEN, J. R. M., HUNTLEY, B. and HASLETT, J. (2015). Bayesian inference for palaeoclimate with time uncertainty and stochastic volatility. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 115–138. [MR3293921](#)
- POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. [MR1723786](#)
- POURAHMADI, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87** 425–435. [MR1782488](#)
- PRISM CLIMATE GROUP, OREGON STATE UNIVERSITY. Available at <http://prism.oregonstate.edu>. Created February 4, 2004.

- R CORE TEAM (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- ROBERTS, G. O. and ROSENTHAL, J. S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist.* **18** 349–367. [MR2749836](#)
- SALTER-TOWNSHEND, M. and HASLETT, J. (2012). Fast inversion of a flexible regression model for multivariate pollen counts data. *Environmetrics* **23** 595–605. [MR3020077](#)
- SAUCEDO-GARCÍA, A., ANAYA, A. L., ESPINOSA-GARCÍA, F. J. and GONZÁLEZ, M. C. (2014). Diversity and communities of foliar endophytic fungi from different agroecosystems of *Coffea arabica* L. in two regions of Veracruz, Mexico. *PLoS ONE* **9** e98454.
- SCHAPIRE, R. E. (1990). The strength of weak learnability. *Mach. Learn.* **5** 197–227.
- SEEGER, M. (2004). Low rank updates for the Cholesky decomposition. Technical Report No. EPFL-REPORT-161468.
- SMITH, M. and KOHN, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *J. Amer. Statist. Assoc.* **97** 1141–1153. [MR1951266](#)
- STAN DEVELOPMENT TEAM (2016). Stan Modeling Language Users Guide and Reference Manual, version 2.10.0 ed. Available at <http://mc-stan.org>.
- STEIN, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer, New York. [MR1697409](#)
- STEIN, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spat. Stat.* **8** 1–19. [MR3326818](#)
- TER BRAAK, C. J. F. and VAN DAM, H. (1989). Inferring pH from diatoms: A comparison of old and new calibration methods. *Hydrobiologia* **178** 209–223.
- TIPTON, J. R., HOOTEN, M. B., NOLAN, C., BOOTH, R. K and MCLACHLAN, J. (2019). Supplement to “Predicting paleoclimate from compositional data using multivariate Gaussian process inverse prediction.” DOI:[10.1214/19-AOAS1281SUPP](https://doi.org/10.1214/19-AOAS1281SUPP).
- VASKO, K., TOIVONEN, H. T. and KORHOLA, A. (2000). A Bayesian multinomial Gaussian response model for organism-based environmental reconstruction. *J. Paleolimnol.* **24** 243–250.
- VECCHIA, A. V. (1988). Estimation and model identification for continuous spatial processes. *J. Roy. Statist. Soc. Ser. B* **50** 297–312. [MR0964183](#)
- WAHL, E. R., DIAZ, H. F. and OHLWEIN, C. (2012). A pollen-based reconstruction of summer temperature in central North America and implications for circulation patterns during medieval times. *Glob. Planet. Change* **84** 66–74.
- WARTON, D. I., BLANCHET, F. G., O’HARA, R. B., OVASKAINEN, O., TASKINEN, S., WALKER, S. C. and HUI, F. K. C. (2015). So many variables: Joint modeling in community ecology. *Trends Ecol. Evol.* **30** 766–779.
- WILLIAMS, J. and SHUMAN, B. (2008). Obtaining accurate and precise environmental reconstructions from the modern analog technique and North American surface pollen dataset. *Quat. Sci. Rev.* **27** 669–687.

PROPENSITY SCORE WEIGHTING FOR CAUSAL INFERENCE WITH MULTIPLE TREATMENTS

BY FAN LI AND FAN LI

Yale University and Duke University

Causal or unconfounded descriptive comparisons between multiple groups are common in observational studies. Motivated from a racial disparity study in health services research, we propose a unified propensity score weighting framework, the balancing weights, for estimating causal effects with multiple treatments. These weights incorporate the generalized propensity scores to balance the weighted covariate distribution of each treatment group, all weighted toward a common prespecified target population. The class of balancing weights include several existing approaches such as the inverse probability weights and trimming weights as special cases. Within this framework, we propose a set of target estimands based on linear contrasts. We further develop the generalized overlap weights, constructed as the product of the inverse probability weights and the harmonic mean of the generalized propensity scores. The generalized overlap weighting scheme corresponds to the target population with the most overlap in covariates across the multiple treatments. These weights are bounded and thus bypass the problem of extreme propensities. We show that the generalized overlap weights minimize the total asymptotic variance of the moment weighting estimators for the pairwise contrasts within the class of balancing weights. We consider two balance check criteria and propose a new sandwich variance estimator for estimating the causal effects with generalized overlap weights. We apply these methods to study the racial disparities in medical expenditure between several racial groups using the 2009 Medical Expenditure Panel Survey (MEPS) data. Simulations were carried out to compare with existing methods.

REFERENCES

- ABADIE, A. and IMBENS, G. W. (2012). A martingale representation for matching estimators. *J. Amer. Statist. Assoc.* **107** 833–843. [MR2980089](#)
- ATHEY, S., IMBENS, G. W. and WAGER, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 597–623. [MR3849336](#)
- BALSA, A. I., CAO, Z. and MCGUIRE, T. G. (2007). Does managed health care reduce health care disparities between minorities and Whites? *J. Health Econ.* **27** 781–807.
- BUNTIN, M. B. and ZASLAVSKY, A. M. (2004). Too much ado about two-part models and transformation? Comparing methods of modeling medicare expenditures. *J. Health Econ.* **23** 525–542.
- CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. [MR3769544](#)

Key words and phrases. Balancing weights, generalized propensity score, generalized overlap weights, health services research, pairwise comparison, racial disparity.

- COOK, B. L., MCGUIRE, T. G. and ZASLAVSKY, A. M. (2012). Measuring racial/ethnic disparities in health care: Methods and practical issues. *Health Serv. Res.* **47** 1232–1254.
- COOK, B. L., MCGUIRE, T. G., MEARA, E. and ZASLAVSKY, A. M. (2009). Adjusting for health status in non-linear models of health care disparities. *Health Serv. Outcomes Res. Methodol.* **9** 1–21.
- COOK, B. L., MCGUIRE, T. G., LOCK, K. and ZASLAVSKY, A. M. (2010). Comparing methods of racial and ethnic disparities measurement across different settings of mental health care. *Health Serv. Res.* **45** 825–847.
- CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. [MR2482144](#)
- DING, P. and LI, F. (2018). Causal inference: A missing data perspective. *Statist. Sci.* **33** 214–237. [MR3797711](#)
- DUDOIT, S. and VAN DER LAAN, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat. Methodol.* **2** 131–154. [MR2161394](#)
- FENG, P., ZHOU, X.-H., ZOU, Q.-M., FAN, M.-Y. and LI, X.-S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Stat. Med.* **31** 681–697. [MR2900870](#)
- HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **1** 25–46.
- HANEUSE, S. and ROTNITZKY, A. (2013). Estimation of the effect of interventions that modify the received treatment. *Stat. Med.* **32** 5260–5277. [MR3141372](#)
- HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. [MR1995826](#)
- HIRSHBERG, D. A. and ZUBIZARRETA, J. R. (2017). On two approaches to weighting in causal inference. *Epidemiology* **28** 812–816.
- IMBENS, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710. [MR1789821](#)
- IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.
- IOM (2003). *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. National Academies Press, Washington, DC.
- JØRGENSEN, B. (1997). *The Theory of Dispersion Models. Monographs on Statistics and Applied Probability* **76**. CRC Press, London. [MR1462891](#)
- LECHNER, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Rev. Econ. Stat.* **84** 205–220.
- LI, L. and GREENE, T. (2013). A weighting analogue to pair matching in propensity score analysis. *Int. J. Biostat.* **9** 215–234. [MR3149627](#)
- LI, F. and LI, F. (2019a). Supplement to “Propensity score weighting for causal inference with multiple treatments.” DOI:[10.1214/19-AOAS1282SUPP](#).
- LI, F. and LI, F. (2019b). Double-robust estimation in difference-in-differences with an application to traffic safety evaluation. *Observational Studies* **5** 1–20.
- LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. [MR3803473](#)
- LI, F., THOMAS, L. E. and LI, F. (2019). Addressing extreme propensity scores via the overlap weights. *Am. J. Epidemiol.* **1** 250–257.
- LI, F., ZASLAVSKY, A. M. and LANDRUM, M. B. (2013). Propensity score weighting with multi-level data. *Stat. Med.* **32** 3373–3387. [MR3074363](#)
- LOPEZ, M. J. and GUTMAN, R. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. *Statist. Sci.* **32** 432–454. [MR3696004](#)
- MANNING, W. G. and MULLAHY, J. (2001). Estimating log models: To transform or not to transform? *J. Health Econ.* **20** 461–494.

- MCCAFFREY, D. F., RIDGEWAY, G. and MORRAL, A. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* **9** 403–425.
- MCCAFFREY, D. F., GRIFFIN, B. A., ALMIRALL, D., SLAUGHTER, M. E., RAMCHAND, R. and BURGETTE, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat. Med.* **32** 3388–3414. [MR3074364](#)
- MCGUIRE, T. G., ALEGRIA, M., COOK, B. L., WELLS, K. B. and ZASLAVSKY, A. M. (2006). Implementing the institute of medicine definition of disparities: An application to mental health care. *Health Serv. Res.* **41** 1979–2005.
- MOORE, K. L., NEUGEBAUER, R., VAN DER LAAN, M. J. and TAGER, I. B. (2012). Causal inference in epidemiological studies with strong confounding. *Stat. Med.* **31** 1380–1404. [MR2925056](#)
- MUÑOZ, I. D. and VAN DER LAAN, M. (2012). Population intervention causal effects based on stochastic interventions. *Biometrics* **68** 541–549. [MR2959621](#)
- PARK, R. (1966). Estimation with heteroscedastic error terms. *Econometrica* **34** 888.
- PIRRACCHIO, R., PETERSEN, M. L. and VAN DER LAAN, M. (2015). Improving propensity score estimators' robustness to model misspecification using super learner. *Am. J. Epidemiol.* **181** 108–119.
- RASSEN, J. A., SHELAT, A. A., FRANKLIN, J. M., GLYNN, R. J., SOLOMON, D. H. and SCHNEEWEISS, S. (2013). Matching by propensity score in cohort studies with three treatment groups. *Epidemiology* **24** 401–409.
- ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. [MR1294730](#)
- ROBINS, J., LI, L., TCHETGEN, E. T. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman* **2** 335–421.
- ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. [MR0742974](#)
- STEFANSKI, L. A. and BOOS, D. D. (2002). The calculus of M -estimation. *Amer. Statist.* **56** 29–38. [MR1939394](#)
- VANDERWEELE, T. J. and ROBINSON, W. R. (2014a). On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* **25** 473–484.
- VANDERWEELE, T. J. and ROBINSON, W. R. (2014b). Rejoinder: How to reduce racial disparities?: Upon what to intervene? *Epidemiology* **25** 491–493.
- VAN DER LAAN, M. J. and PETERSEN, M. L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *Int. J. Biostat.* **3** Art. 3, 54. [MR2306841](#)
- YANG, S., IMBENS, G. W., CUI, Z., FARIES, D. E. and KADZIOLA, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics* **72** 1055–1065. [MR3591590](#)
- YOSHIDA, K., HERNÁNDEZ-DÍAZ, S., SOLOMON, D. H., JACKSON, J. W., GAGNE, J. J., GLYNN, R. J. and FRANKLIN, J. M. (2017). Matching weights to simultaneously compare three treatment groups comparison to three-way matching. *Epidemiology* **28** 387–395.
- ZANUTTO, E., LU, B. and HORNIK, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *J. Educ. Behav. Stat.* **30** 59–73.
- ZASLAVSKY, A. M. and AYANIAN, J. Z. (2005). Integrating research on racial and ethnic disparities in health care over place and time. *Med. Care* **43** 303–307.
- ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* **110** 910–922. [MR3420672](#)

OUTLINE ANALYSES OF THE CALLED STRIKE ZONE IN MAJOR LEAGUE BASEBALL

BY DALE L. ZIMMERMAN, JUN TANG AND RUI HUANG

University of Iowa

We extend statistical shape analytic methods known as outline analysis for application to the strike zone, a central feature of the game of baseball. Although the strike zone is rigorously defined by Major League Baseball's official rules, umpires make mistakes in calling pitches as strikes (and balls) and may even adhere to a strike zone somewhat different than that prescribed by the rule book. Our methods yield inference on geometric attributes (centroid, dimensions, orientation and shape) of this “called strike zone” (CSZ) and on the effects that years, umpires, player attributes, game situation factors and their interactions have on those attributes. The methodology consists of first using kernel discriminant analysis to determine a noisy outline representing the CSZ corresponding to each factor combination, then fitting existing elliptic Fourier and new generalized superelliptic models for closed curves to that outline and finally analyzing the fitted model coefficients using standard methods of regression analysis, factorial analysis of variance and variance component estimation. We apply these methods to PITCHf/x data comprising more than three million called pitches from the 2008–2016 Major League Baseball seasons to address numerous questions about the CSZ. We find that all geometric attributes of the CSZ, except its size, became significantly more like those of the rule-book strike zone from 2008–2016 and that several player attribute/game situation factors had statistically and practically significant effects on many of them. We also establish that the variation in the horizontal center, width and area of an individual umpire's CSZ from pitch to pitch is smaller than their variation among CSZs from different umpires.

REFERENCES

- AHN, S. J. (2004). *Least Squares Orthogonal Distance Fitting of Curves and Surfaces in Space* **3151**. Springer, Berlin.
- BELONGIE, S., MALIK, J. and PUZICHA, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 509–522.
- BLAND, M. (2000). *An Introduction to Medical Statistics*, 3rd ed. Oxford Univ. Press, Oxford.
- BOOKSTEIN, F. L. (1997). *Morphometric Tools for Landmark Data*. Cambridge Univ. Press, Cambridge. [MR1469220](#)
- BROOKS, D., PAVLIDIS, H. and JUDGE, J. (2015). Moving beyond WOWY: A mixed approach to measuring catcher framing. Available at <https://www.baseballprospectus.com/news/article/25514/moving-beyond-wowy-a-mixed-approach-to-measuring-catcher-framing/>.
- CARRUTH, M. (2012). The strike zone. Available at <https://www.lookoutlanding.com/2012/10/29/3561060/the-strike-zone>.

Key words and phrases. Elliptic Fourier model, kernel discriminant analysis, morphometrics, orthogonal distance fitting, shape analysis, superellipse.

- CLAUDE, J. (2008). *Morphometrics with R*. Springer, New York.
- DEL CASTILLO, E. and COLOSIMO, B. M. (2011). Statistical shape analysis of experiments for manufacturing processes. *Technometrics* **53** 1–15. [MR2791943](#)
- DESHPANDE, S. K. and WYNER, A. (2017). A hierarchical Bayesian model of pitch framing. *J. Quant. Anal. Sports* **13** 95–112.
- DRYDEN, I. L. and MARDIA, K. V. (1998). *Statistical Shape Analysis*. *Wiley Series in Probability and Statistics: Probability and Statistics*. Wiley, Chichester. [MR1646114](#)
- DUONG, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *J. Stat. Softw.* **21** 1–16.
- FAST, M. (2011a). Spinning yarn: The real strike zone. Available at <http://www.baseballprospectus.com/article.php?articleid=12965>.
- FAST, M. (2011b). Spinning yarn: The real strike zone, part 2. Available at <http://www.baseballprospectus.com/article.php?articleid=14098>.
- GARDINER, M. (1965). The superellipse: A curve that lies between the ellipse and the rectangle. *Scientific American* **213** 222–234.
- GREEN, E. and DANIELS, D. P. (2014). What does it take to call a strike? Three biases in umpire decision making. Available at http://www.sloansportsconference.com/wp-content/uploads/2014/02/2014_SSAC_What-Does-it-Take-to-Call-a-Strike.pdf.
- GREEN, E. and DANIELS, D. P. (2015). Impact aversion in arbitrator decisions. Available at <https://dx.doi.org/10.2139/ssrn.2391558>.
- GREEN, E. and DANIELS, D. P. (2017). Bayesian instinct. Available at https://site.stanford.edu/sites/default/files/3624-ssrn-id2916929_2.pdf.
- HALL, P. and KANG, K.-H. (2005). Bandwidth choice for nonparametric classification. *Ann. Statist.* **33** 284–306. [MR2157804](#)
- JÄGER, J. M. and SCHÖLLHORN, W. I. (2012). Identifying individuality and variability in team tactics by means of statistical shape analysis and multilayer perceptrons. *Human Movement Science* **31** 303–317.
- KAGAN, D. (2009). The anatomy of a pitch: Doing physics with PITCHf/x data. *The Physics Teacher* **47** 412–416.
- KIM, J. W. and KING, B. G. (2014). Seeing stars: Matthew effects and status bias in Major League Baseball umpiring. *Manage. Sci.* **60** 2619–2644.
- KLASSEN, E., SRIVASTAVA, A., MIO, M. and JOSHI, S. H. (2004). Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 372–383.
- KUHL, F. P. and GIARDINA, C. R. (1982). Elliptic Fourier features of a closed contour. *Computer Graphics and Image Processing* **18** 236–258.
- KURTEK, S., SRIVASTAVA, A., KLASSEN, E. and DING, Z. (2012). Statistical modeling of curves using shapes and related features. *J. Amer. Statist. Assoc.* **107** 1152–1165. [MR3010902](#)
- LENTH, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics* **31** 469–473. [MR1041567](#)
- LINDBERGH, B. (2015). The strike zone time travel test: How much of baseball’s anemic offense really stems from the swollen zone? Available at <http://grantland.com/the-triangle/mlb-larger-strike-zone-declining-offense/>.
- LOPEZ, M. and MILLS, B. (2018). Everyone wants to go home during extra innings—maybe even the umps. Available at <https://fivethirtyeight.com/features/everyone-wants-to-go-home-during-extra-innings-maybe-even-the-umps/>.
- LU, T.-F. C., GRAYBILL, F. A. and BURDICK, R. K. (1987). Confidence intervals on the ratio of expected mean squares $(\theta_1 + d\theta_2)/\theta_3$. *Biometrics* **43** 535–543. [MR0909757](#)
- MARCHI, M. and ALBERT, J. (2014). *Analyzing Baseball Data with R*. CRC Press, Boca Raton, FL.
- MILLS, B. M. (2014). Social pressure at the plate: Inequality aversion, status, and mere exposure. *Managerial and Decision Economics* **35** 387–403.

- MILLS, B. M. (2016a). Policy changes in Major League Baseball: Improved agent behavior and ancillary productivity outcomes. *Economic Inquiry* **55** 1104–1118.
- MILLS, B. M. (2016b). Are the umpires at it again? Available at <https://www.fangraphs.com/tht/are-the-umpires-at-it-again/>.
- MILLS, B. M. (2017). Technological innovations in monitoring and evaluation: Evidence of performance impacts among Major League Baseball umpires. *Labour Econ.* **46** 189–199.
- MOLYNEUX, G. (2016). Prospectus feature: Umpires aren't compassionate; they're Bayesian. Available at <https://www.baseballprospectus.com/news/article/28513/prospectus-feature-umpires-arent-compassionate-theyre-bayesian/>.
- MOSKOWITZ, T. J. and WERTHEIM, L. J. (2011). *Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won*. Random House, New York.
- PARSONS, C. A., SULAEMAN, J., YATES, M. C. and HAMERMESH, D. S. (2011). Strike three: Discrimination, incentives, and evaluation. *Am. Econ. Rev.* **101** 1410–35.
- PASSAN, J. (2015). Sources: MLB could alter strike zone as response to declining offense. Available at <https://sports.yahoo.com/news/sources-mlb-could-alter-strike-zone-as-response-to-declining-offense-232940947.html>.
- ROEGELE, J. (2013a). A simple strike zone formula and calculating plate discipline stats. Available at <https://www.beyondtheboxscore.com/2013/8/5/4576622/simple-strike-zone-formula-calculating-plate-discipline-stats-pitchfx-sabermetrics>.
- ROEGELE, J. (2013b). Investigating the “Lefty Strike.” Available at <https://www.beyondtheboxscore.com/2013/6/7/4391656/investigating-the-lefty-strike-pitchfx-sabermetrics>.
- ROEGELE, J. (2013c). The living strike zone. Available at <https://www.baseballprospectus.com/news/article/21262/baseball-proquestus-the-living-strike-zone/>.
- ROEGELE, J. (2014a). The strike zone during the PITCHf/x era. Available at <https://www.fangraphs.com/tht/the-strike-zone-during-the-pitchfx-era/>.
- ROEGELE, J. (2014b). The strike zone expansion is out of control. Available at <https://www.fangraphs.com/tht/the-strike-zone-expansion-is-out-of-control/>.
- ROEGELE, J. (2015). The 2015 strike zone. Available at <https://www.fangraphs.com/tht/the-2015-strike-zone/>.
- ROEGELE, J. (2016). The 2016 strike zone. Available at <https://www.fangraphs.com/tht/the-2016-strike-zone/>.
- ROSIN, P. L. (2000). Fitting superellipses. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 726–732.
- SIEVERT, C. (2014). Taming PITCHf/x data with XML2R and pitchRx. *R J.* **6** 5–19.
- SIEVERT, C. (2015). pitchRx: Tools for harnessing MLBAM Gameday data and visualizing PITCHf/x. R package version 1.7.
- SRIVASTAVA, A. and KLASSEN, E. P. (2016). *Functional and Shape Data Analysis. Springer Series in Statistics*. Springer, New York. **MR3821566**
- SRIVASTAVA, A., JOSHI, S. H., MIO, W. and LIU, X. (2005). Statistical shape analysis: Clustering, learning, and testing. *IEEE Trans. Pattern Anal. Mach. Intell.* **27** 590–602.
- TAINSKY, S., MILLS, B. M. and WINFREE, J. A. (2015). Further examination of potential discrimination among MLB umpires. *Journal of Sports Economics* **16** 353–374.
- TEUTSCH, C., BERNDT, D., TROSTMANN, E. and WEBER, M. (2013). Real-time detection of elliptic shapes for automated object recognition and object tracking. In *Proceedings of SPIE—The International Society for Optical Engineering*.
- VACCA, J. R. (2007). *Biometric Technologies and Verification Systems*. Elsevier, Amsterdam.
- WALSH, J. (2010). The compassionate umpire. Available at <https://www.fangraphs.com/tht/the-compassionate-umpire/>.
- WAND, M. P. and JONES, M. C. (1994). Multivariate plug-in bandwidth selection. *Comput. Statist.* **9** 97–116. **MR1280754**
- WASSERMAN, L. (2006). *All of Nonparametric Statistics. Springer Texts in Statistics*. Springer, New York. **MR2172729**

- ZELDITCH, M. L., SWIDERSKI, D. L. and SHEETS, H. D. (2012). *Geometric Morphometrics for Biologists: A Primer*, 2nd ed. Elsevier, Amsterdam.
- ZIMMERMAN, D. L., TANG, J. and HUANG, R. (2019). Supplement to “Outline analyses of the called strike zone in Major League Baseball.” DOI:[10.1214/19-AOAS1285SUPPA](https://doi.org/10.1214/19-AOAS1285SUPPA), DOI:[10.1214/19-AOAS1285SUPPB](https://doi.org/10.1214/19-AOAS1285SUPPB), DOI:[10.1214/19-AOAS1285SUPPC](https://doi.org/10.1214/19-AOAS1285SUPPC), DOI:[10.1214/19-AOAS1285SUPPD](https://doi.org/10.1214/19-AOAS1285SUPPD), DOI:[10.1214/19-AOAS1285SUPPE](https://doi.org/10.1214/19-AOAS1285SUPPE), DOI:[10.1214/19-AOAS1285SUPPF](https://doi.org/10.1214/19-AOAS1285SUPPF).

EMPIRICAL BAYES ANALYSIS OF RNA SEQUENCING EXPERIMENTS WITH AUXILIARY INFORMATION

BY KUN LIANG

University of Waterloo

Finding differentially expressed genes is a common task in high-throughput transcriptome studies. While traditional statistical methods rank the genes by their test statistics alone, we analyze an RNA sequencing dataset using the auxiliary information of gene length and the test statistics from a related microarray study. Given the auxiliary information, we propose a novel nonparametric empirical Bayes procedure to estimate the posterior probability of differential expression for each gene. We demonstrate the advantage of our procedure in extensive simulation studies and a psoriasis RNA sequencing study. The companion R package `calm` is available at Bioconductor.

REFERENCES

- ANDREASSEN, O. A., DJUROVIC, S., THOMPSON, W. K., SCHORK, A. J., KENDLER, K. S., O'DONOVAN, M. C., RUJESCU, D., WERGE, T., VAN DE BUNT, M. et al. (2013). Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am. J. Hum. Genet.* **92** 197–209.
- BENIDT, S. and NETTLETON, D. (2015). Simseq: A nonparametric approach to simulation of RNA-sequence datasets. *Bioinformatics* **31** 2131–2140.
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. [MR1325392](#)
- BLANCHARD, G. and ROQUAIN, É. (2009). Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.* **10** 2837–2871. [MR2579914](#)
- BRZYSKI, D., PETERSON, C. B., SOBCZYK, P., CANDÈS, E. J., BOGDAN, M. and SABATTI, C. (2017). Controlling the rate of GWAS false discoveries. *Genetics* **205** 61–75.
- CAI, T. T. and SUN, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.* **104** 1467–1481. [MR2597000](#)
- COLLADO-TORRES, L., NELLORE, A., KAMMERS, K., ELLIS, S. E., TAUB, M. A., HANSEN, K. D., JAFFE, A. E., LANGMEAD, B. and LEEK, J. T. (2017). Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35** 319–321.
- CRAVEN, P. and WAHBA, G. (1978). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–403.
- DU, L. and ZHANG, C. (2014). Single-index modulated multiple testing. *Ann. Statist.* **42** 30–79. [MR3226157](#)
- EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. [MR2054289](#)
- EFRON, B. (2007). Size, power and false discovery rates. *Ann. Statist.* **35** 1351–1377. [MR2351089](#)
- EFRON, B. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Stat.* **2** 197–223. [MR2415600](#)

Key words and phrases. Conditional density, multiple comparison, nonparametric regression, simultaneous inference.

- EFRON, B. (2010). Correlated z -values and the accuracy of large-scale statistical estimates. *J. Amer. Statist. Assoc.* **105** 1042–1055. [MR2752597](#)
- EFRON, B. and TIBSHIRANI, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **23** 70–86.
- EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. [MR1946571](#)
- FAN, J. and HAN, X. (2017). Estimation of the false discovery proportion with unknown dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1143–1164. [MR3689312](#)
- FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. [MR3010887](#)
- FAN, J. and YIM, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika* **91** 819–834. [MR2126035](#)
- FERKINGSTAD, E., FRIGESSI, A., RUE, H., THORLEIFSSON, G. and KONG, A. (2008). Unsupervised empirical Bayesian multiple testing with external covariates. *Ann. Appl. Stat.* **2** 714–735. [MR2524353](#)
- GAGNON-BARTSCH, J. A. and SPEED, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13** 539–552.
- GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 499–517. [MR1924303](#)
- GUDJONSSON, J. E., DING, J., JOHNSTON, A., TEJASVI, T., GUZMAN, A. M., NAIR, R. P., VOORHEES, J. J., ABECASIS, G. R. and ELDER, J. T. (2010). Assessment of the psoriatic transcriptome in a large sample: Additional regulated genes and comparisons with in vitro models. *Journal of Investigative Dermatology* **130** 1829–1840.
- HALL, P., RACINE, J. and LI, Q. (2004). Cross-validation and the estimation of conditional probability densities. *J. Amer. Statist. Assoc.* **99** 1015–1026. [MR2109491](#)
- HUMMEL, M., MEISTER, R. and MANSMANN, U. (2008). GlobalANCOVA: Exploration and assessment of gene group effects. *Bioinformatics* **24** 78–85.
- IGNATIADIS, N., KLAUS, B., ZAUGG, J. B. and HUBER, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods* **13** 577–580.
- JABBARI, A., SUÁREZ-FARIÑAS, M., DEWELL, S. and KRUEGER, J. G. (2012). Transcriptional profiling of psoriasis using RNA-seq reveals previously unidentified differentially expressed genes. *Journal of Investigative Dermatology* **132** 246–249.
- JIN, J. (2008). Proportion of non-zero normal means: Universal oracle equivalences and uniformly consistent estimators. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 461–493. [MR2420411](#)
- JIN, J. and CAI, T. T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506. [MR2325113](#)
- KUKURBA, K. R. and MONTGOMERY, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols* **2015** 951–969.
- LANGAAS, M., LINDQVIST, B. H. and FERKINGSTAD, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 555–572. [MR2168204](#)
- LAW, C. W., CHEN, Y., SHI, W. and SMYTH, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15** R29.
- LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3** e161.
- LI, A. and BARBER, R. F. (2017). Accumulation tests for FDR control in ordered hypothesis testing. *J. Amer. Statist. Assoc.* **112** 837–849. [MR3671774](#)
- LIANG, K. (2019). Supplement to “Empirical Bayes analysis of RNA sequencing experiments with auxiliary information.” DOI:[10.1214/19-AOAS1270SUPP](#).
- 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](#)

- MACDONALD, P., LIANG, K. and JANSSEN, A. (2019). Dynamic adaptive procedures that control the false discovery rate. *Electron. J. Stat.* **13** 3009–3024. [MR4010590](#)
- MARTIN, R. and TOKDAR, S. (2012). A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics* **13** 427–439.
- MEINSHAUSEN, N. and RICE, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **34** 373–393. [MR2275246](#)
- NESTLE, F. O., CONRAD, C., TUN-KYI, A., HOMEY, B., GOMBERT, M., BOYMAN, O., BURG, G., LIU, Y.-J. and GILLIET, M. (2005). Plasmacytoid predendritic cells initiate psoriasis through interferon- α production. *J. Exp. Med.* **202** 135–143.
- NEWTON, M. A. (2002). On a nonparametric recursive estimator of the mixing distribution. *Sankhyā* **64** 306–322. [MR1981761](#)
- OSHLACK, A. and WAKEFIELD, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct* **4** 14.
- PARISI, R., SYMMONS, D. P., GRIFFITHS, C. E., ASHCROFT, D. M. et al. (2013). Global epidemiology of psoriasis: A systematic review of incidence and prevalence. *Journal of Investigative Dermatology* **133** 377–385.
- PATRA, R. K. and SEN, B. (2016). Estimation of a two-component mixture model with applications to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 869–893. [MR3534354](#)
- QU, L., NETTLETON, D. and DEKKERS, J. C. M. (2012). A hierarchical semiparametric model for incorporating intergene information for analysis of genomic data. *Biometrics* **68** 1168–1177. [MR3040023](#)
- ROSENBLATT, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)* 25–31. Academic Press, New York. [MR0254987](#)
- SCHWARTZMAN, A. and LIN, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika* **98** 199–214. [MR2804220](#)
- SCOTT, J. G., KELLY, R. C., SMITH, M. A., ZHOU, P. and KASS, R. E. (2015). False discovery rate regression: An application to neural synchrony detection in primary visual cortex. *J. Amer. Statist. Assoc.* **110** 459–471. [MR3367240](#)
- SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability*. CRC Press, London. [MR0848134](#)
- SMYTH, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3** Art. 3, 29. [MR2101454](#)
- STANGE, J., DICKHAUS, T., NAVARRO, A. and SCHUNK, D. (2016). Multiplicity- and dependency-adjusted p -values for control of the family-wise error rate. *Statist. Probab. Lett.* **111** 32–40. [MR3474779](#)
- STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 187–205. [MR2035766](#)
- SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. [MR2411657](#)
- SWINDELL, W. R., XING, X., VOORHEES, J. J., ELDER, J. T., JOHNSTON, A. and GUDJONSSON, J. E. (2014). Integrative RNA-seq and microarray data analysis reveals GC content and gene length biases in the psoriasis transcriptome. *Physiological Genomics* **46** 533–546.
- TANSEY, W., WANG, Y., BLEI, D. M. and RABADAN, R. (2018). Black box FDR. *International Conference on Machine Learning* 4874–4883.
- TSOI, L. C., IYER, M. K., STUART, P. E., SWINDELL, W. R., GUDJONSSON, J. E., TEJASVI, T., SARKAR, M. K., LI, B., DING, J. et al. (2015). Analysis of long non-coding RNAs highlights tissue-specific expression patterns and epigenetic profiles in normal and psoriatic skin. *Genome Biol.* **16** 24.

- VAN DER FITS, L., VAN DER WEL, L., LAMAN, J. D., PRENS, E. P. and VERSCHUREN, M. C. (2004). In psoriasis lesional skin the type I interferon signaling pathway is activated, whereas interferon- α sensitivity is unaltered. *Journal of Investigative Dermatology* **122** 51–60.
- WANG, J., ZHAO, Q., HASTIE, T. and OWEN, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Ann. Statist.* **45** 1863–1894. [MR3718155](#)
- WASSERMAN, L. (2006). *All of Nonparametric Statistics. Springer Texts in Statistics*. Springer, New York. [MR2172729](#)
- WOOD, S. N. (2003). Thin plate regression splines. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 95–114. [MR1959095](#)
- WOOD, S. N. (2017). *Generalized Additive Models: An Introduction with $\mathbb{B}$ and $\mathbb{R}$* . *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. [MR3726911](#)
- YAO, Y., RICHMAN, L., MOREHOUSE, C., DE LOS REYES, M., HIGGS, B. W., BOUTRIN, A., WHITE, B., COYLE, A., KRUEGER, J. et al. (2008). Type I interferon: Potential therapeutic target for psoriasis? *PLoS ONE* **3** e2737.
- YOUNG, D. S. and HUNTER, D. R. (2010). Mixtures of regressions with predictor-dependent mixing proportions. *Comput. Statist. Data Anal.* **54** 2253–2266. [MR2720486](#)

NEW FORMULATION OF THE LOGISTIC-GAUSSIAN PROCESS TO ANALYZE TRAJECTORY TRACKING DATA¹

BY GIANLUCA MASTRANTONIO^{*,2}, CLARA GRAZIAN^{†,‡,2},
 SARA MANCINELLI[§] AND ENRICO BIBBONA[¶]

Politecnico di Torino^{*}, *Università degli Studi “Gabriele d’Annunzio”*[†],
University of New South Wales[‡], *Università di Roma “La Sapienza”*[§] and
Politecnico di Torino[¶]

Improved communication systems, shrinking battery sizes and the price drop of tracking devices have led to an increasing availability of trajectory tracking data. These data are often analyzed to understand animal behavior.

In this work, we propose a new model for interpreting the animal movement as a mixture of characteristic patterns, that we interpret as different behaviors. The probability that the animal is behaving according to a specific pattern, at each time instant, is nonparametrically estimated using the Logistic-Gaussian process. Owing to a new formalization and the way we specify the coregionalization matrix of the associated multivariate Gaussian process, our model is invariant with respect to the choice of the reference element and of the ordering of the probability vector components. We fit the model under a Bayesian framework, and show that the Markov chain Monte Carlo algorithm we propose is straightforward to implement.

We perform a simulation study with the aim of showing the ability of the estimation procedure to retrieve the model parameters. We also test the performance of the information criterion we used to select the number of behaviors. The model is then applied to a real dataset where a wolf has been observed before and after procreation. The results are easy to interpret, and clear differences emerge in the two phases.

REFERENCES

- AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability*. CRC Press, London. [MR0865647](#)
- ALFREDEÉN, A. C. (2006). Denning behavior and movement pattern during summer of wolves, *Canis lupus*. Grimsö Institute of Wildlife Biology.
- ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. [MR1990662](#)
- ANDERSON, C. R. and LINDZEY, F. G. (2003). Estimating cougar predation rates from GPS location clusters. *J. Wildl. Manag.* **67** 307–316.
- ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* **18** 343–373. [MR2461882](#)
- ANTONACCI, M., BELLOTTI, R., CAFAGNA, F., DE PALMA, M., DIACONO, D., DONVITO, G., ITALIANO, A., GERVASONI, R., MAGGI, G. et al. (2017). The ReCaS Project: The Bari In-

Key words and phrases. Wolf data, invariance, coregionalization, continuous-time hidden Markov model.

- frastructure. In *High Performance Scientific Computing Using Distributed Infrastructures* 17–33 Chapter 3.
- AYRES, L. A., GARDNER, C. L. and FOSTER, J. W. (1991). Den site activity patterns of gray wolves, *Canis lupus*, in Southcentral Alaska. *Can. Field-Nat.* **105** 497–504.
- BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated classification likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 719–725. DOI:10.1109/34.865189.
- BLACKWELL, P. G. (1997). Random diffusion models for animal movement. *Ecol. Model.* **100** 87–102.
- BLACKWELL, P. G. (2003). Bayesian inference for Markov processes with diffusion and discrete components. *Biometrika* **90** 613–627. MR2006839
- BLACKWELL, P. G. (2018). Integrated continuous-time hidden Markov models. arXiv preprint arXiv:1807.11907.
- BLACKWELL, P. G., NIU, M., LAMBERT, M. S. and LAPOINT, S. D. (2016). Exact Bayesian inference for animal movement in continuous time. *Methods Ecol. Evol.* **7** 184–195.
- BRILLINGER, D. R. (2010). Modeling spatial trajectories. In *Handbook of Spatial Statistics. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 463–475. CRC Press, Boca Raton, FL. MR2730961
- BROST, B. M., HOOTEN, M. B., HANKS, E. M. and SMALL, R. J. (2015). Animal movement constraints improve resource selection inference in the presence of telemetry error. *Ecology* **96** 2590–2597.
- BROWNLIE, J. (1911). The mathematical theory of random migration and epidemic distribution. *Proc. Roy. Soc. Edinburgh* **31** 262–289. DOI:10.1017/S0370164600025116.
- BRUNSDON, T. M. and SMITH, T. M. F. (1998). The time series analysis of compositional data. *J. Off. Stat.* **14** 237–253.
- CAGNACCI, F., BOITANI, L., POWELL, R. A. and BOYCE, M. S. (2010). Animal ecology meets GPS-based radiotelemetry: A perfect storm of opportunities and challenges. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **365** 2157–2162. DOI:10.1098/rstb.2010.0107.
- CALABRESE, J. M., FLEMING, C. H. and GURARIE, E. (2016). ctmm: An R package for analyzing animal relocation data as a continuous-time stochastic process. *Methods Ecol. Evol.* **7** 1124–1132.
- CIUCCI, P., BOITANI, L., FRANCISCI, F. and ANDREOLI, G. (1997). Home range, activity and movements of a wolf pack in central Italy. *J. Zool.* **243** 803–819. DOI:10.1111/j.1469-7998.1997.tb01977.x.
- CRESSIE, N. A. C. (1993). *Statistics for Spatial Data. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. Revised reprint of the 1991 edition. MR1239641
- DEMMA, D. J. and MECH, L. D. (2011). Accuracy of estimating wolf summer territories by daytime locations. *Am. Midl. Nat.* **165** 436–445. DOI:10.1674/0003-0031-165.2.436.
- FILZMOSER, P. and HRON, K. (2009). Correlation analysis for compositional data. *Math. Geosci.* **41** 905–919. MR2557238
- FLEMING, C. H. and CALABRESE, J. M. (2019). ctmm: Continuous-time movement modeling. R package version 0.5.3.
- FLEMING, C. H., CALABRESE, J. M., MUELLER, T., OLSON, K. A., LEIMGRUBER, P. and FAGAN, W. F. (2014a). From fine-scale foraging to home ranges: A semivariance approach to identifying movement modes across spatiotemporal scales. *Amer. Nat.* **183** E154–E167.
- FLEMING, C. H., CALABRESE, J. M., MUELLER, T., OLSON, K. A., LEIMGRUBER, P. and FAGAN, W. F. (2014b). Non-Markovian maximum likelihood estimation of autocorrelated movement processes. *Methods Ecol. Evol.* **5** 462–472.
- FRAIR, J. L., FIEBERG, J., HEBBLEWHITE, M., CAGNACCI, F., DECESARE, N. J. and PEDROTTI, L. (2010). Resolving issues of imprecise and habitat-biased locations in ecological analyses using GPS telemetry data. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **365** 2187–2200. DOI:10.1098/rstb.2010.0084.

- FRANKE, A., CAELLI, T., KUZYK, G. and HUDSON, R. J. (2006). Prediction of wolf (*Canis lupus*) kill-sites using hidden Markov models. *Ecol. Model.* **197** 237–246.
- GELFAND, A. E., DIGGLE, P. J., FUENTES, M. and GUTTORP, P., eds. (2010). *Handbook of Spatial Statistics. Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. [MR2761512](#)
- GURARIE, E., FLEMING, C. H., FAGAN, W. F., LAIDRE, K. L., HERNÁNDEZ-PLIEGO, J. and OVASKAINEN, O. (2017). Correlated velocity models as a fundamental unit of animal movement: Synthesis and applications. *Movement Ecology* **5** 13.
- HARRINGTON, F. H. and MECH, L. D. (1982). An analysis of howling response parameters useful for wolf pack censusing. *J. Wildl. Manag.* **46** 686–693.
- HEATON, M. J., DATTA, A., FINLEY, A. O., FURRER, R., GUINNESS, J., GUHANIYOGI, R., GERBER, F., GRAMACY, R. B., HAMMERLING, D. et al. (2019). A case study competition among methods for Analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* **24** 398–425. [MR3996451](#)
- HEBBLEWHITE, M. and MERRILL, E. (2008). Modelling wildlife-human relationships for social species with mixed-effects resource selection models. *J. Appl. Ecol.* **45** 834–844.
- HOOTEN, M. B. and JOHNSON, D. S. (2017). Basis function models for animal movement. *J. Amer. Statist. Assoc.* **112** 578–589. [MR3671753](#)
- HOOTEN, M. B., JOHNSON, D. S., MCCLINTOCK, B. T. and MORALES, J. M. (2017). *Animal Movement: Statistical Models for Telemetry Data*. CRC Press, Boca Raton, FL.
- HOOTEN, M. B., SCHARF, H. R., HEFLEY, T. J., PEARSE, A. T. and WEEGMAN, M. D. (2018). Animal movement models for migratory individuals and groups. *Methods Ecol. Evol.* **9** 1692–1705.
- HOUSTON, A. I. and MCNAMARA, J. M. (1999). *Models of Adaptive Behaviour: An Approach Based on State*. Cambridge Univ. Press, Cambridge.
- JEDRZEJSKI, W., SCHMIDT, K., THEUERKAUF, J., JEDRZEJSKA, B. and OKARMA, H. (2001). Daily movements and territory use by radio-collared wolves (*Canis lupus*) in Białowieża Primeval Forest in Poland. *Can. J. Zool.* **79** 1993–2004. DOI:[10.1139/z01-147](#).
- JOHNSON, D. S., HOOTEN, M. B. and KUHN, C. E. (2013). Estimating animal resource selection from telemetry data using point process models. *J. Anim. Ecol.* **82** 1155–1164.
- JOHNSON, D. S., LONDON, J. M., LEA, M.-A. and DURBAN, J. W. (2008a). Continuous-time correlated random walk model for animal telemetry data. *Ecology* **89** 1208–1215.
- JOHNSON, D. S., THOMAS, D. L., VER HOEF, J. M. and CHRIST, A. (2008b). A general framework for the analysis of animal resource selection from telemetry data. *Biometrics* **64** 968–976. [MR2571498](#)
- JONA LASINIO, G., MASTRANTONIO, G. and POLLICE, A. (2013). Discussing the “big n problem”. *Stat. Methods Appl.* **22** 97–112. [MR3038184](#)
- JONSEN, I. D., FLEMMING, J. M. and MYERS, R. A. (2005). Robust state space modeling of animal movement data. *Ecology* **86** 2874–2880.
- LANGROCK, R., KING, R., MATTHIOPOULOS, J., THOMAS, L., FORTIN, D. and MORALES, J. M. (2012). Flexible and practical modeling of animal telemetry data: Hidden Markov models and extensions. *Ecology* **93** 2336–2342. DOI:[10.1890/11-2241.1](#).
- LANGROCK, R., HOPCRAFT, J. G. C., BLACKWELL, P. G., GOODALL, V., KING, R., NIU, M., PATTERSON, T. A., PEDERSEN, M. W., SKARIN, A. et al. (2014). Modelling group dynamic animal movement. *Methods Ecol. Evol.* **5** 190–199.
- LARK, R. M. and BISHOP, T. F. A. (2007). Cokriging particle size fractions of the soil. *Eur. J. Soil Sci.* **58** 763–774.
- LINDERMAN, S. W., JOHNSON, M. J. and ADAMS, R. P. (2015). Dependent multinomial models made easy: Stick breaking with the Pólya–gamma augmentation. In *Proceedings of the 28th International Conference on Neural Information Processing Systems* 3456–3464.
- LONG, W. and WANG, Q. (2013). Two methods of correlation coefficient on compositional data. *Proc. Comput. Sci.* **18** 1757–1763. 2013 International Conference on Computational Science.

- MANCINELLI, S., BOITANI, L. and CIUCCI, P. (2018). Determinants of home range size and space use patterns in a protected wolf (*Canis lupus*) population in the central Apennines, Italy. *Can. J. Zool.* **96** 828–838.
- MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics. Wiley Series in Probability and Statistics*. Wiley, Chichester. Revised reprint of *Statistics of Directional Data* by Mardia [MR0336854 (49 #1627)]. [MR1828667](#)
- MARTINS, A. B. T., BONAT, W. H. and RIBEIRO, P. J. JR. (2016). Likelihood analysis for a class of spatial geostatistical compositional models. *Spat. Stat.* **17** 121–130. [MR3532093](#)
- MARUOTTI, A., PUNZO, A., MASTRANTONIO, G. and LAGONA, F. (2015). A time-dependent extension of the projected normal regression model for longitudinal circular data based on a hidden Markov heterogeneity structure. *Stoch. Environ. Res. Risk Assess.* **30** 1725–1740.
- MASTRANTONIO, G., JONA LASINIO, G. and GELFAND, A. E. (2016). Spatio-temporal circular models with non-separable covariance structure. *TEST* **25** 331–350. [MR3493522](#)
- MASTRANTONIO, G., MARUOTTI, A. and JONA-LASINIO, G. (2015). Bayesian hidden Markov modelling using circular-linear general projected normal distribution. *Environmetrics* **26** 145–158. [MR3324908](#)
- MCCLINTOCK, B. T., KING, R., THOMAS, L., MATTHIOPOULOS, J., MCCONNELL, B. J. and MORALES, J. M. (2012). A general discrete-time modeling framework for animal movement using multistate random walks. *Ecol. Monogr.* **82** 335–349. DOI:[10.1890/11-0326.1](#).
- MECH, D. and BOITANI, L. (2003). *Wolves: Behavior, Ecology and Conservation*. Univ. Chicago Press, Chicago, IL.
- MECH, L. D., WOLF, P. C. and PACKARD, J. M. (1999). Regurgitative food transfer among wild wolves. *Can. J. Zool.* **77** 1192–1195.
- MERRILL, S. B. and DAVID MECH, L. (2000). Details of extensive movements by Minnesota wolves (*Canis lupus*). *Am. Midl. Nat.* **144** 428–433.
- MORALES, J. M., HAYDON, D. T., FRAIR, J., HOLSINGER, K. E. and FRYXELL, J. M. (2004). Extracting more out of relocation data: Building movement models as mixtures of random walks. *Ecology* **85** 2436–2445.
- MULLER, P. (2017). *Bayesian Nonparametric Data Analysis*. CreateSpace Independent Publishing Platform.
- NIU, M., BLACKWELL, P. G. and SKARIN, A. (2016). Modeling interdependent animal movement in continuous time. *Biometrics* **72** 315–324. [MR3515758](#)
- PACI, L. and FINAZZI, F. (2018). Dynamic model-based clustering for spatio-temporal data. *Stat. Comput.* **28** 359–374. [MR3747568](#)
- PARTON, A. and BLACKWELL, P. G. (2017). Bayesian inference for multistate ‘step and turn’ animal movement in continuous time. *J. Agric. Biol. Environ. Stat.* **22** 373–392. [MR3692470](#)
- PATTERSON, T. A., BASSON, M., BRAVINGTON, M. V. and GUNN, J. S. (2009). Classifying movement behaviour in relation to environmental conditions using hidden Markov models. *J. Anim. Ecol.* **78** 1113–1123.
- PATTERSON, T. A., PARTON, A., LANGROCK, R., BLACKWELL, P. G., THOMAS, L. and KING, R. (2017). Statistical modelling of individual animal movement: An overview of key methods and a discussion of practical challenges. *ASTA Adv. Stat. Anal.* **101** 399–438. [MR3712406](#)
- PAWLOWSKY, V. and BURGER, H. (1992). Spatial structure analysis of regionalized compositions. *Math. Geol.* **24** 675–691.
- PETERS, R. P. (1979). Mental maps in wolf territoriality. In *The Behavior and Ecology of Wolves* (E. K. Klinghammer, ed.) 119–152. Garland STPM Press, New York.
- POHLE, J., LANGROCK, R., VAN BEEST, F. M. and SCHMIDT, N. M. (2017). Selecting the number of states in hidden Markov models: Pragmatic solutions illustrated using animal movement. *J. Agric. Biol. Environ. Stat.* **22** 270–293. [MR3692465](#)
- 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](#)

- ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika* **97** 539–550. [MR2672482](#)
- ROTHMAN, R. J. and MECH, L. D. (1979). Scent-marking in lone wolves and newly formed pairs. *Anim. Behav.* **27** 750–760.
- SAND, H., ZIMMERMANN, B., WABAKKEN, P., ANDREN, H. and PEDERSEN, H. C. (2010). Using GPS technology and GIS cluster analyses to estimate kill rates in wolf-ungulate ecosystems. *Wildl. Soc. Bull.* **33** 914–925.
- SCHARF, H. R., HOOTEN, M. B., JOHNSON, D. S. and DURBAN, J. W. (2018). Process convolution approaches for modeling interacting trajectories. *Environmetrics* **29** e2487, 18. [MR3799912](#)
- THEUERKAUF, J. (2009). What drives wolves: Fear or hunger? Humans, diet, climate and wolf activity patterns. *Ethology* **115** 649–657.
- TJELMELAND, H. and LUND, K. V. (2003). Bayesian modelling of spatial compositional data. *J. Appl. Stat.* **30** 87–100. [MR1957363](#)
- VILA, C. (1995). Observations on the daily activity patterns in the Iberian wolf. *Ecology and Conservation of Wolves in a Changing World* **35** 335–340.
- WANG, F. and GELFAND, A. E. (2013). Directional data analysis under the general projected normal distribution. *Stat. Methodol.* **10** 113–127. [MR2974815](#)
- ZUCCHINI, W. and MACDONALD, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R. Monographs on Statistics and Applied Probability* **110**. CRC Press, Boca Raton, FL. [MR2523850](#)

A SIMPLE, CONSISTENT ESTIMATOR OF SNP HERITABILITY FROM GENOME-WIDE ASSOCIATION STUDIES¹

BY ARMIN SCHWARTZMAN*, ANDREW J. SCHORK[†], RONG ZABLOCKI* AND
WESLEY K. THOMPSON*,[†]

*University of California, San Diego** and *Institute of Biological Psychiatry*[†]

Analysis of genome-wide association studies (GWAS) is characterized by a large number of univariate regressions where a quantitative trait is regressed on hundreds of thousands to millions of single-nucleotide polymorphism (SNP) allele counts, one at a time. This article proposes an estimator of the SNP heritability of the trait, defined here as the fraction of the variance of the trait explained by the SNPs in the study. The proposed GWAS heritability (GWASH) estimator is easy to compute, highly interpretable and is consistent as the number of SNPs and the sample size increase. More importantly, it can be computed from summary statistics typically reported in GWAS, not requiring access to the original data. The estimator takes full account of the linkage disequilibrium (LD) or correlation between the SNPs in the study through moments of the LD matrix, estimable from auxiliary datasets. Unlike other proposed estimators in the literature, we establish the theoretical properties of the GWASH estimator and obtain analytical estimates of the precision, allowing for power and sample size calculations for SNP heritability estimates and forming a firm foundation for future methodological development.

REFERENCES

- THE 1000 GENOMES PROJECT CONSORTIUM, AUTON, A., BROOKS, L. D., DURBIN, R. M., GARRISON, E. P., KANG, H. M., KORBEL, J. O., MARCHINI, J. L., MCCARTHY, S. et al. (2015). A global reference for human genetic variation. *Nature* **526** 68–74.
- BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. [MR2387969](#)
- BULIK-SULLIVAN, B. K., LOH, P. R., FINUCANE, H. K., RIPKE, S., YANG, J., SCHIZOPHRENIA WORKING GROUP OF THE PSYCHIATRIC GENOMICS CONSORTIUM, PATTERSON, N., DALY, M. J., PRICE, A. L. et al. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47** 291–295.
- CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. [MR2676885](#)
- CHANG, C. C., CHOW, C. C., TELLIER, L. C., VATTIKUTI, S., PURCELL, S. M. and LEE, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Giga-science* **4** 7.
- DE LEEUW, C. A., MOOIJ, J. M., HESKES, T. and POSTHUMA, D. (2015). MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11** e1004219.
- DICKER, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika* **101** 269–284. [MR3215347](#)

Key words and phrases. High dimensional data, massively univariate regression, summary statistics, single nucleotide polymorphism.

- FALCONER, D. S. and MACKAY, T. F. C. (1996). *Introduction to quantitative genetics*, 4th ed. Longman, Harlow.
- FISHER, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52** 399–433.
- GAZAL, S., FINUCANE, H. K., FURLOTTE, N. A., LOH, P.-R., PALAMARA, P. F., LIU, X., SCHOECH, A., BULIK-SULLIVAN, B., NEALE, B. M. et al. (2017). Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49** 1421.
- HEMANI, G., SHAKHBAZOV, K., WESTRA, H.-J., ESKO, T., HENDERS, A. K., MCRAE, A. F., YANG, J., GIBSON, G., MARTIN, N. G. et al. (2014a). Detection and replication of epistasis influencing transcription in humans. *Nature* **508** 249.
- HEMANI, G., SHAKHBAZOV, K., WESTRA, H.-J., ESKO, T., HENDERS, A. K., MCRAE, A. F., YANG, J., GIBSON, G., MARTIN, N. G. et al. (2014b). Another explanation for apparent epistasis. *Nature* **514** E5.
- HILL, W. G., GODDARD, M. E. and VISSCHER, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4** e1000008.
- HOFERT, M., KOJADINOVIC, I., MAECHLER, M. and YAN, J. (2014). *copula: Multivariate dependence with copulas*. R package version 0.999-9.
- KOJADINOVIC, I., JUN YAN, J. Y. et al. (2010). Modeling multivariate distributions with continuous margins using the copula R package. *J. Stat. Softw.* **34** 1–20.
- LI, Y., WILLER, C., SANNA, S. and ABECASIS, G. (2009). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Genotype Imputation* **10** 387–406.
- LOCKE, A. E. et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518** 197–206.
- LYNCH, M. and WALSH, B. (1998). *Genetics and Analysis of Quantitative Traits*. Vol. 1. Sinauer Sunderland, MA.
- MACARTHUR, J., BOWLER, E., CEREZO, M., GIL, L., HALL, P., HASTINGS, E., JUNKINS, H., MCMAHON, A., MILANO, A. et al. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res.* **45** D896–D901.
- OKBAY, A. et al. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533** 539–542.
- PASANIUC, B. and PRICE, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18** 117–127.
- SCHORK, A. J., THOMPSON, W. K., PHAM, P., TORKAMANI, A., RODDEY, J. C., SULLIVAN, P. F., KELSOE, J. R., O'DONOVAN, M. C., FURBERG, H. et al. (2013). All SNPs are not created equal: Genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet.* **9** e1003449.
- SCHWARTZMAN, A., SCHORK, A. J., ZABLOCKI, R. and THOMPSON, W. K. (2019). Supplement to “A simple, consistent estimator of SNP heritability from genome-wide association studies.” DOI:10.1214/19-AOAS1291SUPPA, DOI:10.1214/19-AOAS1291SUPPB.
- SNIEKERS, S., STRINGER, S., WATANABE, K., JANSEN, P. R., COLEMAN, J. R. I., KRAPOHL, E., TASKESSEN, E., HAMMERSCHLAG, A. R., OKBAY, A. et al. (2017). Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49** 1107–1112.
- SPAIN, S. L. and BARRETT, J. C. (2015). Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* **24** R111–R119.
- SPEED, D., CAI, N., CONSORTIUM, U., JOHNSON, M. R., NEJENTSEV, S. and BALDING, D. J. (2017). Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49** 986–992.
- VISSCHER, P. M., WRAY, N. R., ZHANG, Q., SKLAR, P., MCCARTHY, M. I., BROWN, M. A. and YANG, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **101** 5–22.

- WOOD, A. R., TUKE, M. A., NALLS, M. A., HERNANDEZ, D. G., BANDINELLI, S., SINGLETON, A. B., MELZER, D., FERRUCCI, L., FRAYLING, T. M. et al. (2014). Another explanation for apparent epistasis. *Nature* **514** E3–E5.
- YANG, J., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G. et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42** 565–569.
- YANG, J., BAKSHI, A., ZHU, Z., HEMANI, G., VINKHUYZEN, A. A., LEE, S. H., ROBINSON, M. R., PERRY, J. R., NOLTE, I. M. et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47** 1114.
- ZHOU, X., CARBONETTO, P. and STEPHENS, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* **9** e1003264.

A HIERARCHICAL CURVE-BASED APPROACH TO THE ANALYSIS OF MANIFOLD DATA

BY LIBERTY VITTERT*, ADRIAN W. BOWMAN*,¹ AND STANISLAV KATINA^{†,1}

*University of Glasgow** and *Masaryk University[†]*

One of the data structures generated by medical imaging technology is high resolution point clouds representing anatomical surfaces. Stereophotogrammetry and laser scanning are two widely available sources of this kind of data. A standardised surface representation is required to provide a meaningful correspondence across different images as a basis for statistical analysis. Point locations with anatomical definitions, referred to as landmarks, have been the traditional approach. Landmarks can also be taken as the starting point for more general surface representations, often using templates which are warped on to an observed surface by matching landmark positions and subsequent local adjustment of the surface.

The aim of the present paper is to provide a new approach which places anatomical curves at the heart of the surface representation and its analysis. Curves provide intermediate structures which capture the principal features of the manifold (surface) of interest through its ridges and valleys. As landmarks are often available these are used as anchoring points, but surface curvature information is the principal guide in estimating the curve locations. The surface patches between these curves are relatively flat and can be represented in a standardised manner by appropriate surface transects to give a complete surface model.

This new approach does not require the use of a template, reference sample or any external information to guide the method and, when compared with a surface based approach, the estimation of curves is shown to have improved performance. In addition, examples involving applications to mussel shells and human faces show that the analysis of curve information can deliver more targeted and effective insight than the use of full surface information.

REFERENCES

- ARMANN, R. and BALTHOFF, I. (2012). Male and female faces are only perceived categorically when linked to familiar identities. And when in doubt, he is a male. *Vis. Res.* **63** 69–80.
- BOOKSTEIN, F. L. (1997). Landmark methods for forms without landmarks: Morphometrics of group differences in outline shape. *Med. Image Anal.* **1** 225–243.
- BOWMAN, A. W. and AZZALINI, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, Oxford.
- BOWMAN, A. W., KATINA, S., SMITH, J. and BROWN, D. (2015). Anatomical curve identification. *Comput. Statist. Data Anal.* **86** 52–64. [MR3312737](#)
- BRUCE, V., BURTON, A. M., HANNA, E., HEALEY, P., MASON, O., COOMBES, A., FRIGHT, R. and LINNEY, A. (1993). Sex discrimination: How do we tell the difference between male and female faces? *Perception* **23** 131–152.

Key words and phrases. Anatomy, curves, manifold, p-splines, shape analysis, smoothing, three dimensional.

- CHE, W., ZHANG, X., ZHANG, Y.-K., PAUL, J.-C. and XU, B. (2011). Ridge extraction of a smooth 2-manifold surface based on vector field. *Comput. Aided Geom. Design* **28** 215–232. [MR2802876](#)
- CHENG, W., DRYDEN, I. L. and HUANG, X. (2016). Bayesian registration of functions and curves. *Bayesian Anal.* **11** 447–475. [MR3471998](#)
- CLAES, P., WALTERS, M., SHRIVER, M. D., PUTS, D., GIBSON, G., CLEMENT, J., BAYNAM, G., VERBEKE, G., VANDERMEULEN, D. et al. (2012). Sexual dimorphism in multiple aspects of 3d facial symmetry and asymmetry defined by spatially dense geometric morphometrics. *J. Anat.* **221** 97–114.
- COHEN-STEINER, D., DE VERDIERE, E. C. and YVINEC, M. (2002). Conforming Delaunay triangulations in 3D. In *Proceedings of the Eighteenth Annual Symposium on Computational Geometry* 199–208. ACM, New York.
- DAVIES, R. H., COOTES, T. F. and TAYLOR, C. J. (2001). A minimum description length approach to statistical shape modelling. In *Biennial International Conference on Information Processing in Medical Imaging* 50–63. Springer.
- DE BERG, M., VAN KREVELD, M., OVERMARS, M. and SCHWARZKOPF, O. (2000). *Computational Geometry: Algorithms and Applications*, revised ed. Springer, Berlin. [MR1763734](#)
- DRYDEN, I. L. and MARDIA, K. V. (2016). *Statistical Shape Analysis*, 2nd ed. Wiley, New York. [MR1646114](#)
- EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with *B*-splines and penalties. *Statist. Sci.* **11** 89–121. [MR1435485](#)
- FARKAS, L. (1994). *Anthropometry of the Head and Face*, 2nd ed. Raven Press, New York.
- FITZER, S. C., VITTERT, L., BOWMAN, A., KAMENOS, N. A., PHOENIX, V. R. and CUSACK, M. (2015). Ocean acidification and temperature increase impact mussel shell shape and thickness: Problematic for protection? *Ecol. Evol.* **5** 4875–4884.
- GOLDFEATHER, J. and INTERRANTE, V. (2004). A novel cubic-order algorithm for approximating principal direction vectors. *ACM Trans. Graph.* **23** 45–63.
- HAMMOND, P., HUTTON, T. J., ALLANSON, J. E., CAMPBELL, L. E., HENNEKAM, R., HOLDEN, S., PATTON, M. A., SHAW, A., TEMPLE, I. K. et al. (2004). 3d analysis of facial morphology. *Am. J. Med. Genet., Part A* **126** 339–348.
- HASTIE, T. and STUETZLE, W. (1989). Principal curves. *J. Amer. Statist. Assoc.* **84** 502–516. [MR1010339](#)
- KATINA, S., MCNEIL, K., AYOUB, A., GUILFOYLE, B., KHAMBAY, B., SIEBERT, P., SUKNO, F., ROJAS, M., VITTERT, L. et al. (2016). The definitions of three-dimensional landmarks on the human face: An interdisciplinary view. *J. Anat.* **228** 355–365.
- KENT, J. T., MARDIA, K. V. and WEST, J. (1996). Ridge curves and shape analysis. In *BMVC Proceedings* 1996 1–10.
- KNEIP, A. and RAMSAY, J. O. (2008). Combining registration and fitting for functional models. *J. Amer. Statist. Assoc.* **103** 1155–1165. [MR2528838](#)
- KOENDERINK, J. J. (1990). *Solid Shape. MIT Press Series in Artificial Intelligence*. MIT Press, Cambridge, MA. [MR1045203](#)
- KOENDERINK, J. and VAN DOORN, A. (1992). Surface shape and curvature scales. *Image Vis. Comput.* **10** 557–564.
- MAO, Z., JU, X., SIEBERT, J. P., COCKSHOT, W. P. and AYOUB, A. (2006). Constructing dense correspondences for the analysis of 3d facial morphology. *Pattern Recogn. Lett.* **27** 597–608.
- MEIER, D. and FISHER, E. (2002). Parameter space warping: Shape-based correspondence between morphologically different objects. *IEEE Trans. Med. Imag.* **21** 31–47.
- MEYER, M., BARR, A., LEE, H. and DESBRUN, M. (2002). Generalized barycentric coordinates on irregular polygons. *J. Graphics Tools* **7** 13–22.
- OHTAKE, Y., BELYAIEV, A. and SEIDEL, H.-P. (2004). Ridge-valley lines on meshes via implicit surface fitting. In *ACM Transactions on Graphics (TOG)* **23** 609–612. ACM.

- PATRANGENARU, V. and ELLINGSON, L. (2016). *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*. CRC Press, Boca Raton, FL. [MR3444169](#)
- PAULSEN, R. R. and HILGER, K. B. (2003). Shape modelling using Markov random field restoration of point correspondences. In *Biennial International Conference on Information Processing in Medical Imaging* 1–12. Springer, Berlin.
- PENNEC, X., AYACHE, N. and THIRION, J.-P. (2000). Landmark-based registration using features identified through differential geometry. In *Handbook of Medical Imaging—Processing and Analysis* 499–513. Academic Press.
- RAVIV, D., BAYRO-CORROCHANO, E. and RASKAR, R. (2017). LRA: Local rigid averaging of stretchable non-rigid shapes. *Int. J. Comput. Vis.* **124** 132–144. [MR3681619](#)
- RAVIV, D. and KIMMEL, R. (2015). Affine invariant geometry for non-rigid shapes. *Int. J. Comput. Vis.* **111** 1–11. [MR3300516](#)
- RAVIV, D., BRONSTEIN, A. M., BRONSTEIN, M. M., WAISMAN, D., SOCHEN, N. and KIMMEL, R. (2014). Equi-affine invariant geometry for shape analysis. *J. Math. Imaging Vision* **50** 144–163. [MR3233140](#)
- ROHR, K. (2001). *Landmark-Based Image Analysis: Using Geometric and Intensity Models* **21**. Springer, Berlin.
- RUSTAMOV, R. M., OVSJANIKOV, M., AZENCOT, O., BEN-CHEN, M., CHAZAL, F. and GUIBAS, L. (2013). Map-based exploration of intrinsic shape differences and variability. *ACM Trans. Graph.* **32** 72.
- SRIVASTAVA, A., SAMIR, C., JOSHI, S. H. and DAOUDI, M. (2009). Elastic shape models for face analysis using curvilinear coordinates. *J. Math. Imaging Vision* **33** 253–265. [MR2480989](#)
- STYLIANOU, G. and FARIN, G. (2004). Crest lines for surface segmentation and flattening. *IEEE Trans. Vis. Comput. Graph.* **5** 536–544.
- VITTERT, L., BOWMAN, A. and KATINA, S. (2019a). Supplement A to “A hierarchical curve-based approach to the analysis of manifold data.” DOI:[10.1214/19-AOAS1267SUPPA](#).
- VITTERT, L., BOWMAN, A. and KATINA, S. (2019b). Supplement B to “A hierarchical curve-based approach to the analysis of manifold data.” DOI:[10.1214/19-AOAS1267SUPPB](#).
- VITTERT, L., BOWMAN, A. and KATINA, S. (2019c). Supplement C to “A hierarchical curve-based approach to the analysis of manifold data.” DOI:[10.1214/19-AOAS1267SUPPC](#).
- VITTERT, L., BOWMAN, A. and KATINA, S. (2019d). Supplement D to “A hierarchical curve-based approach to the analysis of manifold data.” DOI:[10.1214/19-AOAS1267SUPPD](#).
- VITTERT, L., BOWMAN, A. and KATINA, S. (2019e). Supplement E to “A hierarchical curve-based approach to the analysis of manifold data.” DOI:[10.1214/19-AOAS1267SUPPE](#).

SCALABLE HIGH-RESOLUTION FORECASTING OF SPARSE SPATIOTEMPORAL EVENTS WITH KERNEL METHODS: A WINNING SOLUTION TO THE NIJ “REAL-TIME CRIME FORECASTING CHALLENGE”

BY SETH FLAXMAN^{*,1}, MICHAEL CHIRICO[†], PAU PEREIRA[‡] AND CHARLES LOEFFLER[§]

Imperial College London^{}, Grab[†], Amazon, Inc.[‡] and University of Pennsylvania[§]*

We propose a generic spatiotemporal event forecasting method which we developed for the National Institute of Justice’s (NIJ) Real-Time Crime Forecasting Challenge (National Institute of Justice (2017)). Our method is a spatiotemporal forecasting model combining scalable randomized Reproducing Kernel Hilbert Space (RKHS) methods for approximating Gaussian processes with autoregressive smoothing kernels in a regularized supervised learning framework. While the smoothing kernels capture the two main approaches in current use in the field of crime forecasting, kernel density estimation (KDE) and self-exciting point process (SEPP) models, the RKHS component of the model can be understood as an approximation to the popular log-Gaussian Cox Process model. For inference, we discretize the spatiotemporal point pattern and learn a log-intensity function using the Poisson likelihood and highly efficient gradient-based optimization methods. Model hyperparameters including quality of RKHS approximation, spatial and temporal kernel lengthscales, number of autoregressive lags and bandwidths for smoothing kernels as well as cell shape, size and rotation, were learned using cross validation. Resulting predictions significantly exceeded baseline KDE estimates and SEPP models for sparse events.

REFERENCES

- ADAMS, R. P., MURRAY, I. and MACKAY, D. J. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning* 9–16. ACM, New York.
- ADEPEJU, M., ROSSER, G. and CHENG, T. (2016). Novel evaluation metrics for sparse spatiotemporal point process hotspot predictions—A crime case study. *Int. J. Geogr. Inf. Sci.* **30** 2133–2154.
- BERK, R., HEIDARI, H., JABBARI, S., KEARNS, M. and ROTH, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* 0049124118782533.
- BHATT, S., CAMERON, E., FLAXMAN, S. R., WEISS, D. J., SMITH, D. L. and GETHING, P. W. (2017). Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. *J. R. Soc. Interface* **14** 20170520.
- BRIX, A. and DIGGLE, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 823–841. [MR1872069](#)

Key words and phrases. Spatial statistics, time series, supervised learning, spatiotemporal forecasting, Cox process, RKHS.

- CAPLAN, J. M., KENNEDY, L. W. and MILLER, J. (2011). Risk terrain modeling: Brokering criminological theory and GIS methods for crime forecasting. *Justice Q.* **28** 360–381.
- CHANEY, S. P. (2013). Examining the influence of cell size and bandwidth size on kernel density estimation crime hotspot maps for predicting spatial patterns of crime. *Bull. Geogr. Soc. Liege* **60** 7–19.
- CHANEY, S. and RATCLIFFE, J. (2005). *GIS and Crime Mapping*. Wiley, New York.
- CHANEY, S., TOMPSON, L. and UHLIG, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Secur. J.* **21** 4–28.
- COHEN, J., GORR, W. L. and OLLIGSCHLAEGER, A. M. (2007). Leading indicators and spatial interactions: A crime-forecasting model for proactive police deployment. *Geogr. Anal.* **39** 105–127.
- CORBETT-DAVIES, S., PIERSON, E., FELLER, A., GOEL, S. and HUQ, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 797–806. ACM, New York.
- CUNNINGHAM, J. P., SHENOY, K. V. and SAHANI, M. (2008). Fast Gaussian process methods for point process intensity estimation. In *Proceedings of the 25th International Conference on Machine Learning* 192–199. ACM, New York.
- DIGGLE, P. J., MORAGA, P., ROWLINGSON, B. and TAYLOR, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm. *Statist. Sci.* **28** 542–563. [MR3161587](#)
- DRESSEL, J. and FARID, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **4** eaao5580.
- FLAXMAN, S. R. (2014). A general approach to prediction and forecasting crime rates with Gaussian processes. Technical report, Heinz College of Information Systems and Public Policy, Carnegie Mellon Univ., Pittsburgh, PA.
- FLAXMAN, S., WILSON, A., NEILL, D., NICKISCH, H. and SMOLA, A. (2015). Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. In *International Conference on Machine Learning* 607–616.
- FLAXMAN, S., CHIRICO, M., PEREIRA, P. and LOEFFLER, C. (2019a). Supplement to “Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ ‘Real-Time Crime Forecasting Challenge’.” DOI:[10.1214/19-AOAS1284SUPPA](#).
- FLAXMAN, S., CHIRICO, M., PEREIRA, P. and LOEFFLER, C. (2019b). Source code for “Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ ‘Real-Time Crime Forecasting Challenge’.” DOI:[10.1214/19-AOAS1284SUPPB](#).
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.
- GERBER, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decis. Support Syst.* **61** 115–125.
- GORR, W. L. (2009). Forecast accuracy measures for exception reporting using receiver operating characteristic curves. *Int. J. Forecast.* **25** 48–61.
- GORR, W. L. and LEE, Y. (2015). Early warning system for temporary crime hot spots. *J. Quant. Criminol.* **31** 25–47.
- GORR, W., OLLIGSCHLAEGER, A. and THOMPSON, Y. (2003). Short-term forecasting of crime. *Int. J. Forecast.* **19** 579–594.
- GROFF, E. and TANIGUCHI, T. (2019). Using citizen notification to interrupt near-repeat residential burglary patterns: The micro-level near-repeat experiment. *J. Exp. Criminol.* **15** 115–149.
- GUTTORP, P. and GNEITING, T. (2005). On the Whittle–Matérn correlation family. National Research Center for Statistics and the Environment-Technical Report Series, Seattle, WA.
- HEATON, M. J., DATTA, A., FINLEY, A. O., FURRER, R., GUINNESS, J., GUHANYOGI, R., GERBER, F., GRAMACY, R. B., HAMMERLING, D. et al. (2019). A case study competition among methods for analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* **24** 398–425.

- HENNIG, P., OSBORNE, M. A. and GIROLAMI, M. (2015). Probabilistic numerics and uncertainty in computations. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **471** 20150142, 17. [MR3378744](#)
- HUNT, J. M. (2016). Do crime hot spots move? Exploring the effects of the modifiable areal unit problem and modifiable temporal unit problem on crime hot spot stability. Ph.D. thesis, American Univ., Washington, DC.
- JOHNSON, S. D., BOWERS, K. J., BIRKS, D. J. and PEASE, K. (2009). Predictive mapping of crime by ProMap: Accuracy, units of analysis, and the environmental backcloth. In *Putting Crime in Its Place* (D. Weisburd, W. Bernasco and G. Bruinsma, eds.) 165–192. Springer, Dordrecht.
- KANG, H.-W. and KANG, H.-B. (2017). Prediction of crime occurrence from multi-modal data using deep learning. *PLoS ONE* **12** e0176244.
- LEVINE, N. (2004). CrimeStat: A spatial statistics program for the analysis of crime incident locations, version 3.0. Technical report, Ned Levine and Associates/National Institute of Justice, Washington, DC.
- LIU, H. and BROWN, D. E. (2003). Criminal incident prediction using a point-pattern-based density model. *Int. J. Forecast.* **19** 603–622.
- LLOYD, C., GUNTER, T., OSBORNE, M. and ROBERTS, S. (2015). Variational inference for Gaussian process modulated Poisson processes. In *International Conference on Machine Learning* 1814–1822.
- LOEFFLER, C. and FLAXMAN, S. (2018). Is gun violence contagious? A spatiotemporal test. *J. Quant. Criminol.* **34** 999–1017.
- LUM, K. and ISAAC, W. (2016). To predict and serve? *Significance* **13** 14–19.
- MAKRIDAKIS, S., SPILIOTIS, E. and ASSIMAKOPOULOS, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE* **13** e0194889.
- MAY, A., BAGHERI GARAKANI, A., LU, Z. et al. (2019). Kernel approximation methods for speech recognition. *J. Mach. Learn. Res.* **20** Paper No. 59, 36. [MR3960913](#)
- MILTON, P., COUPLAND, H. GIORGI, E. and BHATT, S. (2019). Spatial analysis made easy with linear regression and kernels. *Epidemics*. DOI:[10.1016/j.epidem.2019.100362](#).
- MITCHELL, S., POTASH, E., BAROCAS, S., D'AMOUR, A. and LUM, K. (2018). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. Available at [arXiv:1811.07867](#).
- MOHLER, G. (2013). Modeling and estimation of multi-source clustering in crime and security data. *Ann. Appl. Stat.* **7** 1525–1539. [MR3127957](#)
- MOHLER, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast.* **30** 491–497.
- MOHLER, G. and PORTER, M. D. (2018). Rotational grid, PAI-maximizing crime forecasts. *Stat. Anal. Data Min.* **11** 227–236. [MR3859025](#)
- MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. [MR2816705](#)
- MOHLER, G. O., SHORT, M. B., MALINOWSKI, S., JOHNSON, M., TITA, G. E., BERTOZZI, A. L. and BRANTINGHAM, P. J. (2015). Randomized controlled field trials of predictive policing. *J. Amer. Statist. Assoc.* **110** 1399–1411. [MR3449035](#)
- MØLLER, J. and RASMUSSEN, J. G. (2005). Perfect simulation of Hawkes processes. *Adv. in Appl. Probab.* **37** 629–646. [MR2156552](#)
- MØLLER, J., SYVERVEEN, A. R. and WAAGEPETERSEN, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Stat.* **25** 451–482. [MR1650019](#)
- O'HAGAN, A. (1992). Some Bayesian numerical analysis. In *Bayesian Statistics, 4* (Peñíscola, 1991) 345–363. Oxford Univ. Press, New York. [MR1380285](#)
- NATIONAL INSTITUTE OF JUSTICE (2017). Real-time crime forecasting challenge. Available at <http://www.nij.gov/funding/Pages/fy16-crime-forecasting-challenge.aspx>.

- OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.
- PEASE, K. et al. (1998). *Repeat Victimization: Taking Stock* **90**. Home Office Police Research Group, London.
- PERRY, W. L., MCINNIS, B., PRICE, C. C., SMITH, S. C. and HOLLYWOOD, J. S. (2013). Predictive policing: The role of crime forecasting in law enforcement operations. Technical report, RAND Corporation, Santa Monica, CA.
- PORTER, M. D. and REICH, B. J. (2012). Evaluating temporally weighted kernel density methods for predicting the next event location in a series. *Ann. GIS* **18** 225–240.
- RAHIMI, A. and RECHT, B. (2007). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems* 1177–1184.
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR2514435](#)
- RODRIGUES, A. and DIGGLE, P. J. (2012). Bayesian estimation and prediction for inhomogeneous spatiotemporal log-Gaussian Cox processes using low-rank models, with application to criminal surveillance. *J. Amer. Statist. Assoc.* **107** 93–101. [MR2949344](#)
- ROSSER, G. and CHENG, T. (2019). Improving the robustness and accuracy of crime prediction with the self-exciting point process through isotropic triggering. *Appl. Spatial Anal. Policy* **12** 5–25.
- ROSSER, G., DAVIES, T., BOWERS, K. J., JOHNSON, S. D. and CHENG, T. (2016). Predictive crime mapping: Arbitrary grids or street networks? *J. Quant. Criminol.* **33** 569–594.
- RUDIN, C. and USTUN, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces* **48** 449–466.
- SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Cambridge, MA.
- SCHUTT, H. G. (1922). Advanced police methods in Berkeley. *Natl. Munic. Rev.* **11** 80–85.
- SHIROTA, S. and GELFAND, A. E. (2017). Space and circular time log Gaussian Cox processes with application to crime event data. *Ann. Appl. Stat.* **11** 481–503. [MR3693535](#)
- SNOEK, J., LAROCHELLE, H. and ADAMS, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems* 2951–2959.
- SRIPERUMBUDUR, B. K., FUKUMIZU, K. and LANCKRIET, G. R. G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.* **12** 2389–2410. [MR2825431](#)
- SUN, Y., LI, B. and GENTON, M. G. (2012). Geostatistics for large datasets. In *Advances and Challenges in Space-Time Modelling of Natural Events* 55–77. Springer, New York.
- TADDY, M. A. (2010). Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime. *J. Amer. Statist. Assoc.* **105** 1403–1417. [MR2796559](#)
- TEH, Y. W. and RAO, V. (2011). Gaussian process modulated renewal processes. In *Advances in Neural Information Processing Systems* 2474–2482.
- WANG, X., GERBER, M. S. and BROWN, D. E. (2012). Automatic crime prediction using events extracted from Twitter posts. In *Social Computing Behavioral—Cultural Modeling and Prediction* 231–238. Springer, Berlin.
- WEINBERGER, K., DASGUPTA, A., LANGFORD, J., SMOLA, A. and ATTENBERG, J. (2009). Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning. ICML '09* 1113–1120. ACM, New York.
- 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](#)

ON BAYESIAN NEW EDGE PREDICTION AND ANOMALY DETECTION IN COMPUTER NETWORKS¹

BY SILVIA METELLI* AND NICHOLAS HEARD*,[†]

*Imperial College London** and *Heilbronn Institute for Mathematical Research[†]*

Monitoring computer network traffic for anomalous behaviour presents an important security challenge. Arrivals of new edges in a network graph represent connections between a client and server pair not previously observed, and in rare cases these might suggest the presence of intruders or malicious implants. We propose a Bayesian model and anomaly detection method for simultaneously characterising existing network structure and modelling likely new edge formation. The method is demonstrated on real computer network authentication data and successfully identifies some machines which are known to be compromised.

REFERENCES

- CAHILL, M. H., LAMBERT, D., PINHEIRO, J. C. and SUN, D. X. (2002). Detecting fraud in the real world. In *Handbook of Massive Data Sets* 911–929. Kluwer Academic, Dordrecht.
- CHO, H., DHILLON, I. S., GUAN, Y. and SRA, S. (2004). Minimum sum-squared residue co-clustering of gene expression data. In *Proceedings of the Fourth SIAM International Conference on Data Mining* 114–125. SIAM, Philadelphia, PA. [MR2388433](#)
- COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. [MR0341758](#)
- DHILLON, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 269–274. ACM, New York.
- FISHER, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.
- GHAHRAMANI, Z., GRIFFITHS, T. L. and SOLLICH, P. (2007). Bayesian nonparametric latent feature models. In *Bayesian Statistics 8. Oxford Sci. Publ.* 201–226. Oxford Univ. Press, Oxford. [MR2433194](#)
- HALL, E. C. and WILLETT, R. M. (2016). Tracking dynamic point processes on networks. *IEEE Trans. Inform. Theory* **62** 4327–4346. [MR3515754](#)
- HEARD, N. and METELLI, S. (2014). Modelling new edge formation in a computer network through Bayesian variable selection. In *Joint Intelligence and Security Informatics Conference (JISIC)*, 2014 *European* 272–275. IEEE, New York.
- HEARD, N. and METELLI, S. (2016). Model-based clustering and new edge modelling in a large computer network. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2016 91–96. IEEE, New York.
- HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. [MR0718088](#)
- KENT, A. D. (2015a). Comprehensive, multi-source cyber-security events. Los Alamos National Laboratory, Washington, DC.

Key words and phrases. Computer networks, anomaly detection, new edges, Bayesian inference.

- KENT, A. D. (2015b). Cybersecurity data sources for dynamic network research. In *Dynamic Networks in Cybersecurity* Imperial College Press, London.
- LEE, M., SHEN, H., HUANG, J. Z. and MARRON, J. S. (2010). Biclustering via sparse singular value decomposition. *Biometrics* **66** 1087–1095. [MR2758496](#)
- LI, S., XIE, Y., FARAJTABAR, M., VERMA, A. and SONG, L. (2017). Detecting changes in dynamic events over networks. *IEEE Trans. Signal Inform. Process. Netw.* **3** 346–359. [MR3661705](#)
- LINDERMAN, S. W. and ADAMS, R. P. (2014). Discovering latent network structure in point process data. In *Proceedings of the 31st International Conference on Machine Learning* 1413–1421.
- MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. [MR2758523](#)
- METELLI, S. and HEARD, N. (2019). Supplement to “On Bayesian new edge prediction and anomaly detection in computer networks.” DOI:[10.1214/19-AOAS1286SUPP](#).
- NEIL, J., HASH, C., BRUGH, A., FISK, M. and STORLIE, C. B. (2013). Scan statistics for the online detection of locally anomalous subgraphs. *Technometrics* **55** 403–414. [MR3176546](#)
- PATCHA, A. and PARK, J. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Netw.* **51** 3448–3470.
- PEARSON, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika* **25** 379–410.
- PERRY, P. O. and WOLFE, P. J. (2013). Point process modelling for directed interaction networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 821–849. [MR3124793](#)
- ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. [MR2893856](#)
- RUBIN-DELANCHY, P., PRIEBE, C. E., TANG, M. and CAPE, J. (2017). A statistical interpretation of spectral embedding: The generalised random dot product graph. Preprint. Available at [arXiv:1709.05506](#).
- SCHOLZ, F.-W. and STEPHENS, M. A. (1987). k -sample Anderson–Darling tests. *J. Amer. Statist. Assoc.* **82** 918–924. [MR0910001](#)
- SILL, M., KAISER, S., BENNER, A. and KOPP-SCHNEIDER, A. (2011). Robust biclustering by sparse singular value decomposition incorporating stability selection. *Bioinformatics* **27** 2089–2097.
- SNIJEDERS, T., VAN DE BUNT, G. and STEGLICH, C. (2010). Introduction to stochastic actor-based models for network dynamics. *Soc. Netw.* **32** 44–60.
- SOLOMON, H. and STEPHENS, M. (1978). Approximations to density functions using Pearson curves. *J. Amer. Statist. Assoc.* **73** 153–160.
- SUSSMAN, D. L., TANG, M., FISHKIND, D. E. and PRIEBE, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* **107** 1119–1128. [MR3010899](#)
- TADDY, M. A. (2010). Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime. *J. Amer. Statist. Assoc.* **105** 1403–1417. [MR2796559](#)
- TURCOTTE, M. J., HEARD, N. A. and NEIL, J. (2014). Detecting localised anomalous behaviour in a computer network. In *Advances in Intelligent Data Analysis XIII—13th International Symposium* 321–332. Springer, Berlin.
- ZAMMIT-MANGION, A., ALAN DEWAR, M., KADIRKAMANATHAN, V. and SANGUINETTI, G. (2012). Point process modelling of the Afghan War Diary. *Proc. Natl. Acad. Sci. USA* **109** 12414–12419.
- ZHOU, K., ZHA, H. and SONG, L. (2013). Learning triggering kernels for multi-dimensional Hawkes processes. In *Proceedings of the 30th International Conference on Machine Learning* **28** 1301–1309.

BAYESIAN INDICATOR VARIABLE SELECTION TO INCORPORATE HIERARCHICAL OVERLAPPING GROUP STRUCTURE IN MULTI-OMICS APPLICATIONS

BY LI ZHU^{1,*}, ZHIGUANG HUO[†], TIANZHOU MA^{1,‡}, STEFFI OESTERREICH^{*}
AND GEORGE C. TSENG^{1,*}

University of Pittsburgh^{}, University of Florida[†] and University of Maryland[‡]*

Variable selection is a pervasive problem in modern high-dimensional data analysis where the number of features often exceeds the sample size (a.k.a. small-n-large-p problem). Incorporation of group structure knowledge to improve variable selection has been widely studied. Here, we consider prior knowledge of a hierarchical overlapping group structure to improve variable selection in regression setting. In genomics applications, for instance, a biological pathway contains tens to hundreds of genes and a gene can be mapped to multiple experimentally measured features (such as its mRNA expression, copy number variation and methylation levels of possibly multiple sites). In addition to the hierarchical structure, the groups at the same level may overlap (e.g., two pathways can share common genes). Incorporating such hierarchical overlapping groups in traditional penalized regression setting remains a difficult optimization problem. Alternatively, we propose a Bayesian indicator model that can elegantly serve the purpose. We evaluate the model in simulations and two breast cancer examples, and demonstrate its superior performance over existing models. The result not only enhances prediction accuracy but also improves variable selection and model interpretation that lead to deeper biological insight of the disease.

REFERENCES

- ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. [MR1224394](#)
- BATES, D. and EDELBUETTTEL, D. (2013). Fast and elegant numerical linear algebra using the RcppEigen package. *J. Stat. Softw.* **52** 1–24.
- CHEN, R.-B., CHU, C.-H., YUAN, S. and WU, Y. N. (2016). Bayesian sparse group selection. *J. Comput. Graph. Statist.* **25** 665–683. [MR3533632](#)
- CIRIELLO, G., GATZA, M. L., BECK, A. H., WILKERSON, M. D., RHIE, S. K., PASTORE, A., ZHANG, H., MCLELLAN, M., YAU, C. et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163** 506–519.
- EBERLY, L. E. and CARLIN, B. P. (2000). Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Stat. Med.* **19** 2279–2294.
- GELFAND, A. E. and SAHU, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *J. Amer. Statist. Assoc.* **94** 247–253. [MR1689229](#)
- GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

Key words and phrases. Bayesian variable selection, hierarchical overlapping group structure, overlapping groups, spike and slab.

- GEWEKE, J. et al. (1991). Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN.
- HERNÁNDEZ-LOBATO, D., HERNÁNDEZ-LOBATO, J. M. and DUPONT, P. (2013). Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. *J. Mach. Learn. Res.* **14** 1891–1945. [MR3104499](#)
- JACOB, L., OBOZINSKI, G. and VERT, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning* 433–440. ACM, New York.
- JENATTON, R., MAIRAL, J., OBOZINSKI, G. and BACH, F. (2011). Proximal methods for hierarchical sparse coding. *J. Mach. Learn. Res.* **12** 2297–2334. [MR2825428](#)
- JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. [MR2089135](#)
- KUO, L. and MALICK, B. (1998). Variable selection for regression models. *Sankhya, Ser. B* **60** 65–81. [MR1717076](#)
- KYUNG, M., GILL, J., GHOSH, M. and CASELLA, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal.* **5** 369–411. [MR2719657](#)
- LIU, J., JI, S., YE, J. et al. (2009). SLEP: Sparse learning with efficient projections. The Biodesign Institute, Arizona State University, Tempe, AZ.
- LOCK, E. F. and DUNSON, D. B. (2017). Bayesian genome- and epigenome-wide association studies with gene level dependence. *Biometrics* **73** 1018–1028. [MR3713135](#)
- MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. [MR0997578](#)
- NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.
- O’HARA, R. B. and SILLANPÄÄ, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Anal.* **4** 85–117. [MR2486240](#)
- PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. [MR2524001](#)
- RAVINDRANATH, A. and CADIGAN, K. (2016). The role of the C-clamp in wnt-related colorectal cancers. *Cancers* **8** 74.
- RICHARDSON, S., TSENG, G. C. and SUN, W. (2016). Statistical methods in integrative genomics. *Annu. Rev. Stat. Appl.* **3** 181–209.
- SANTIAGO, L., DANIELS, G., WANG, D., DENG, F.-M. and LEE, P. (2017). Wnt signaling pathway protein LEF1 in cancer, as a biomarker for prognosis and a target for treatment. *American Journal of Cancer Research* **7** 1389.
- SIKORA, M. J., JACOBSEN, B. M., LEVINE, K., CHEN, J., DAVIDSON, N. E., LEE, A. V., ALEXANDER, C. M. and OESTERREICH, S. (2016). WNT4 mediates estrogen receptor signaling and endocrine resistance in invasive lobular carcinoma cell lines. *Breast Cancer Res.* **18** 92.
- SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. [MR3173712](#)
- STINGO, F. C., CHEN, Y. A., TADESSE, M. G. and VANNUCCI, M. (2011). Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Ann. Appl. Stat.* **5** 1978–2002. [MR2884929](#)
- TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. [MR0898357](#)
- TASDEMIR, N., BOSSART, E. A., LI, Z., ZHU, L., SIKORA, M. J., LEVINE, K. M., JACOBSEN, B. M., TSENG, G. C., DAVIDSON, N. E. et al. (2018). Comprehensive phenotypic characterization of human invasive lobular carcinoma cell lines in 2D and 3D cultures. *Cancer Res.* 6209–6222.

- TEO, K., GÓMEZ-CUADRADO, L., TENHAGEN, M., BYRON, A., RÄTZE, M., VAN AMERSFOORT, M., RENES, J., STRENGMAN, E., MANDOLI, A. et al. (2018). E-cadherin loss induces targetable autocrine activation of growth factor signalling in lobular breast cancer. *Sci. Rep.* **8** 15454.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- TURASHVILI, G., BOUCHAL, J., BAUMFORTH, K., WEI, W., DZIECHCIARKOVA, M., EHLMANN, J., KLEIN, J., FRIDMAN, E., SKARDA, J. et al. (2007). Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer* **7** 55.
- VAN HENGEL, J., VANHOENACKER, P., STAES, K. and VAN ROY, F. (1999). Nuclear localization of the p120ctn armadillo-like catenin is counteracted by a nuclear export signal and by E-cadherin expression. *Proc. Natl. Acad. Sci. USA* **96** 7980–7985.
- XU, X. and GHOSH, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Anal.* **10** 909–936. [MR3432244](#)
- 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](#)
- ZHANG, L., BALADANDAYUTHAPANI, V., MALLICK, B. K., MANYAM, G. C., THOMPSON, P. A., BONDY, M. L. and DO, K.-A. (2014a). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 595–620. [MR3258055](#)
- ZHANG, L., MORRIS, J. S., ZHANG, J., ORLOWSKI, R. Z. and BALADANDAYUTHAPANI, V. (2014b). Bayesian joint selection of genes and pathways: Applications in multiple myeloma genomics. *Cancer Inform.* **13** 113.
- ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497. [MR2549566](#)
- ZHU, L., HUO, Z., MA, T., OESTERREICH, S. and TSENG, G. C. (2019). Supplement to “Bayesian indicator variable selection to incorporate hierarchical overlapping group structure in multi-omics applications.” DOI:[10.1214/19-AOAS1271SUPP](#).
- 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](#)

HIERARCHICAL INFINITE FACTOR MODELS FOR IMPROVING THE PREDICTION OF SURGICAL COMPLICATIONS FOR GERIATRIC PATIENTS

BY ELIZABETH LORENZI*, RICARDO HENAO[†] AND KATHERINE HELLER[†]

Berry Consultants and Duke University[†]*

Nearly a third of all surgeries performed in the United States occur for patients over the age of 65; these older adults experience a higher rate of post-operative morbidity and mortality. To improve the care for these patients, we aim to identify and characterize high risk geriatric patients to send to a specialized perioperative clinic while leveraging the overall surgical population to improve learning. To this end, we develop a hierarchical infinite latent factor model (HIFM) to appropriately account for the covariance structure across subpopulations in data. We propose a novel Hierarchical Dirichlet Process shrinkage prior on the loadings matrix that flexibly captures the underlying structure of our data while sharing information across subpopulations to improve inference and prediction. The stick-breaking construction of the prior assumes an infinite number of factors and allows for each subpopulation to utilize different subsets of the factor space and select the number of factors needed to best explain the variation. We develop the model into a latent factor regression method that excels at prediction and inference of regression coefficients. Simulations validate this strong performance compared to baseline methods. We apply this work to the problem of predicting surgical complications using electronic health record data for geriatric patients and all surgical patients at Duke University Health System (DUHS). The motivating application demonstrates the improved predictive performance when using HIFM in both area under the ROC curve and area under the PR Curve while providing interpretable coefficients that may lead to actionable interventions.

REFERENCES

- AHRQ (2016). Healthcare cost and utilization project (hcup) surgery flag software. <https://www.hcup-us.ahrq.gov/toolsoftware/surgflags/surgeryflags.jsp>.
- ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. [MR1224394](#)
- AVALOS-PACHECO, A., ROSSELL, D. and SAVAGE, R. S. (2018). Heterogeneous large datasets integration using Bayesian factor regression. Preprint. Available at [arXiv:1810.09894](#).
- BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. [MR2806429](#)
- CARON, F. and DOUCET, A. (2008). Sparse Bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine Learning* 88–95. ACM, New York.

Key words and phrases. Bayesian factor model, nonparametrics, transfer learning, hierarchical modeling, surgical outcomes, health care.

- CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008a). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. [MR2655722](#)
- CHEN, M., SILVA, J., PAISLEY, J., WANG, C., DUNSON, D. and CARIN, L. (2010). Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Trans. Signal Process.* **58** 6140–6155. [MR2790088](#)
- COREY, K. M., KASHYAP, S., LORENZI, E., LAGOO-DEENADAYALAN, S. A., HELLER, K., WHALEN, K., BALU, S., HEFLIN, M. T., McDONALD, S. R. et al. (2018). Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study. *PLoS Med.* **15** e1002701.
- DESEBBE, O., LANZ, T., KAIN, Z. and CANNESON, M. (2016). The perioperative surgical home: An innovative, patient-centred and cost-effective perioperative care model. *Anaesth. Crit. Care Pain Med.* **35** 59–66.
- ELIXHAUSER, A., STEINER, C., HARRIS, D. R. and RM, C. (1998). Comorbidity measures for use with administrative data. *Med. Care* **36**.
- ETZIONI, D. A., LIU, J. H., O’CONNELL, J. B., MAGGARD, M. A. and KO, C. Y. (2003). Elderly patients in surgical workloads: A population-based analysis. *Am. J. Surg.* **69** 961–965.
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. [MR0350949](#)
- GONG, J. J., SUNDT, T. M., RAWN, J. D. and GUTTAG, J. V. (2015). Instance weighting for patient-specific risk stratification models. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 369–378. ACM, New York.
- HANOVER, N. (2001). Operative mortality with elective surgery in older adults. *Eff. Clin. Pract.* **4** 172–177.
- HEALEY, M. A., SHACKFORD, S. R., OSLER, T. M., ROGERS, F. B. and BURNS, E. (2002). Complications in surgical patients. *Arch. Surg.* **137** 611–618.
- ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. [MR1952729](#)
- ISHWARAN, H. and JAMES, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information. *J. Comput. Graph. Statist.* **11** 508–532. [MR1938445](#)
- JONES, T. S., DUNN, C. L., WU, D. S., CLEVELAND, J. C., KILE, D. and ROBINSON, T. N. (2013). Relationship between asking an older adult about falls and surgical outcomes. *J. Am. Med. Assoc. Surg.* **148** 1132–1138.
- LEE, G., RUBINFELD, I. and SYED, Z. (2012). Adapting surgical models to individual hospitals using transfer learning. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on* 57–63. IEEE, New York.
- LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14** 41–67. [MR2036762](#)
- LORENZI, E., HENAO, R. and HELLER, K. (2019). Supplement to “Hierarchical infinite factor models for improving the prediction of surgical complications for geriatric patients.” DOI:[10.1214/19-AOAS1292SUPPA](#), DOI:[10.1214/19-AOAS1292SUPPB](#), DOI:[10.1214/19-AOAS1292SUPPC](#).
- LUCAS, J., CARVALHO, C., WANG, Q., BILD, A., NEVINS, J. and WEST, M. (2006). Sparse statistical modelling in gene expression genomics. *Bayesian Inference for Gene Expression and Proteomics* **1** 1.
- MCDONALD, S. R., HEFLIN, M. T., WHITSON, H. E., DALTON, T. O., LIDSKY, M. E., LIU, P., POER, C. M., SLOANE, R., THACKER, J. K. et al. (2018). Association of integrated care coordination with postsurgical outcomes in high-risk older adults: The perioperative optimization of senior health (POSH) initiative. *JAMA Surg.* DOI: [10.1001/jamasurg.2017.5513](#)

- MCPARLAND, D., GORMLEY, I. C., MCCORMICK, T. H., CLARK, S. J., KABUDULA, C. W. and COLLINSON, M. A. (2014). Clustering South African households based on their asset status using latent variable models. *Ann. Appl. Stat.* **8** 747–776. [MR3262533](#)
- MCPARLAND, D., PHILLIPS, C. M., BRENNAN, L., ROCHE, H. M. and GORMLEY, I. C. (2017). Clustering high-dimensional mixed data to uncover sub-phenotypes: Joint analysis of phenotypic and genotypic data. *Stat. Med.* **36** 4548–4569. [MR3731239](#)
- MURPHY, K., GORMLEY, I. C. and VIROLI, C. (2017). Infinite mixtures of infinite factor analysers: Nonparametric model-based clustering via latent gaussian models. Preprint. Available at [arXiv:1701.07010](#).
- NI, Y., MUELLER, P. and JI, Y. (2018). Bayesian double feature allocation for phenotyping with electronic health records. Preprint. Available at [arXiv:1809.08988](#).
- POLSON, N. G. and SCOTT, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics 9* 501–538. Oxford Univ. Press, Oxford. [MR3204017](#)
- RAVAL, M. V. and ESKANDARI, M. K. (2012). Outcomes of elective abdominal aortic aneurysm repair among the elderly: Endovascular versus open repair. *Surgery* **151** 245–260.
- ROČKOVÁ, V. and GEORGE, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *J. Amer. Statist. Assoc.* **111** 1608–1622. [MR3601721](#)
- SEO, D. M., GOLDSCHMIDT-CLERMONT, P. J. and WEST, M. (2007). Of mice and men: Space statistical modeling in cardiovascular genomics. *Ann. Appl. Stat.* **1** 152–178. [MR2393845](#)
- SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. [MR1309433](#)
- SPEZIALE, G., NASSO, G., BARATTONI, M. C., ESPOSITO, G., POPOFF, G., ARGANO, V., GRECO, E., SCORCIN, M., ZUSSA, C. et al. (2011). Short-term and long-term results of cardiac surgery in elderly and very elderly patients. *J. Thorac. Cardiovasc. Surg.* **141** 725–731.
- 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](#)
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- WEST, M. (2003). Bayesian factor regression models in the “large p , small n ” paradigm. In *Bayesian Statistics, 7 (Tenerife, 2002)* 733–742. Oxford Univ. Press, New York. [MR2003537](#)
- WIENS, J., GUTTAG, J. and HORVITZ, E. (2014). A study in transfer learning: Leveraging data from multiple hospitals to enhance hospital-specific predictions. *J. Am. Med. Inform. Assoc.* **21** 699–706.
- 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](#)

A GENERAL THEORY FOR PREFERENTIAL SAMPLING IN ENVIRONMENTAL NETWORKS¹

BY JOE WATSON*, JAMES V. ZIDEK* AND GAVIN SHADDICK[†]

*University of British Columbia** and *University of Exeter[†]*

This paper presents a general model framework for detecting the preferential sampling of environmental monitors recording an environmental process across space and/or time. This is achieved by considering the joint distribution of an environmental process with a site-selection process that considers where and when sites are placed to measure the process. The environmental process may be spatial, temporal or spatio-temporal in nature. By sharing random effects between the two processes, the joint model is able to establish whether site placement was stochastically dependent of the environmental process under study. Furthermore, if stochastic dependence is identified between the two processes, then inferences about the probability distribution of the spatio-temporal process will change, as will predictions made of the process across space and time. The embedding into a spatio-temporal framework also allows for the modelling of the dynamic site-selection process itself. Real-world factors affecting both the size and location of the network can be easily modelled and quantified. Depending upon the choice of the population of locations considered for selection across space and time under the site-selection process, different insights about the precise nature of preferential sampling can be obtained. The general framework developed in the paper is designed to be easily and quickly fit using the R-INLA package. We apply this framework to a case study involving particulate air pollution over the UK where a major reduction in the size of a monitoring network through time occurred. It is demonstrated that a significant response-biased reduction in the air quality monitoring network occurred, namely the relocation of monitoring sites to locations with the highest pollution levels, and the routine removal of sites at locations with the lowest. We also show that the network was consistently unrepresenting levels of particulate matter seen across much of GB throughout the operating life of the network. Finally we show that this may have led to a severe overreporting of the population-average exposure levels experienced across GB. This could have great impacts on estimates of the health effects of black smoke levels.

REFERENCES

- AINSLIE, B., REUTEN, C., STEYN, D. G., LE, N. D. and ZIDEK, J. V. (2009). Application of an entropy-based Bayesian optimization technique to the redesign of an existing monitoring network for single air pollutants. *J. Environ. Manag.* **90** 2715–2729.
- BADDELEY, A., RUBAK, E. and TURNER, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. Chapman & Hall/CRC, Boca Raton, FL.

Key words and phrases. Preferential sampling, INLA, random fields, mobile monitors, health effects, air pollution, big data.

- CHANG, H., FU, A. Q., LE, N. D. and ZIDEK, J. V. (2007). Designing environmental monitoring networks to measure extremes. *Environ. Ecol. Stat.* **14** 301–321. [MR2405332](#)
- COLLS, J. (2002). *Air Pollution, Modelling, and Mitigation*. Routledge, Abingdon.
- CONN, P. B., THORSON, J. T. and JOHNSON, D. S. (2017). Confronting preferential sampling when analyzing population distributions: Diagnosis and model-based triage. *Methods Ecol. Evol.* **8** 1535–1545.
- DIGGLE, P. J., MENEZES, R. and SU, T. (2010). Geostatistical inference under preferential sampling. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59** 191–232. [MR2744471](#)
- EPA (2005). Air quality criteria for ozone and related photochemical oxidants. Technical report. Available at <http://oaspub.epa.gov/eims/eimsapi.dispdetail?deid=149923>.
- FITHIAN, W. and HASTIE, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *Ann. Appl. Stat.* **7** 1917–1939. [MR3161707](#)
- FUGLSTAD, G.-A., SIMPSON, D., LINDGREN, F. and RUE, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *J. Amer. Statist. Assoc.* **114** 445–452. [MR3941267](#)
- GELFAND, A. E., SAHU, S. K. and HOLLAND, D. M. (2012). On the effect of preferential sampling in spatial prediction. *Environmetrics* **23** 565–578. [MR3020075](#)
- GOULD, A. L., BOYE, M. E., CROWTHER, M. J., IBRAHIM, J. G., QUARTEY, G., MICALLEF, S. and BOIS, F. Y. (2015). Joint modeling of survival and longitudinal non-survival data: Current methods and issues. Report of the DIA Bayesian joint modeling working group. *Stat. Med.* **34** 2181–2195. [MR3354152](#)
- ISAACS, E. H. and SRIVASTAVA, R. M. (1988). Spatial continuity measures for probabilistic and deterministic geostatistics. *Math. Geol.* **20** 313–341. [MR0950455](#)
- LI, Q. and SU, L. (2018). Accommodating informative dropout and death: A joint modelling approach for longitudinal and semicompeting risks data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 145–163. [MR3758759](#)
- LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. [MR2853727](#)
- LOPERFIDO, N. and GUTTORP, P. (2008). Network bias in air quality monitoring design. *Environmetrics* **19** 661–671. [MR2530390](#)
- MCMILLAN, R. and MURPHY, J. (2017). Measuring the effects of severe air pollution: Evidence from the UK clean air act.
- MICALCOVÁ, D., LVONČIK, S., CHYTRÝ, M. and HÁJEK, O. (2011). Bias in vegetation databases? A comparison of stratified-random and preferential sampling. *J. Veg. Sci.* **22** 281–291.
- OLEA, R. A. (2007). Declustering of clustered preferential sampling for histogram and semivariogram inference. *Math. Geol.* **39** 453–467.
- PATI, D., REICH, B. J. and DUNSON, D. B. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika* **98** 35–48. [MR2804208](#)
- REIS, S., LISKA, T., STEINLE, S., CARNELL, E., LEAVER, D., ROBERTS, E., VIENO, M., BECK, R. and DRAGOSITS, U. (2017). Uk gridded population 2011 based on census 2011 and land cover map 2015. NERC Environmental Information Data Centre. Available at <https://doi.org/10.5285/0995e94d-6d42-40c1-8ed4-5090d82471e1>.
- RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. [MR2649602](#)
- RUE, H., RIEBLER, A., SØRBYE, S. H., ILLIAN, J. B., SIMPSON, D. P. and LINDGREN, F. K. (2017). Bayesian computing with INLA: A review. *Annual Rev. Stat. Appl.* **4** 395–421.
- SCHUMACHER, P. and ZIDEK, J. V. (1993). Using prior information in designing intervention detection experiments. *Ann. Statist.* **21** 447–463. [MR1212187](#)

- SHADDICK, G. and ZIDEK, J. V. (2014). A case study in preferential sampling: Long term monitoring of air pollution in the UK. *Spat. Stat.* **9** 51–65. [MR3326831](#)
- 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](#)
- SWITZER, P. (1977). Estimation of spatial distributions from point sources with application to air pollution measurement. Technical Report 9, Dept. Statistics, Stanford Univ., Stanford, CA.
- WARTON, D. I. and SHEPHERD, L. C. (2010). Poisson point process models solve the “pseudo-absence problem” for presence-only data in ecology. *Ann. Appl. Stat.* **4** 1383–1402. [MR2758333](#)
- WATSON, J. (2019). Towards a general theory for preferential sampling: Extending the log-Gaussian cox process approach to spatio-temporal applications. In preparation.
- WATSON, J., ZIDEK, J. V. and SHADDICK, G. (2019). Supplement to “A general theory for preferential sampling in environmental networks.” DOI:[10.1214/19-AOAS1288SUPP](#).
- WU, L. (2009). *Mixed Effects Models for Complex Data. Monographs on Statistics and Applied Probability* **113**. CRC Press, Boca Raton, FL. [MR2598844](#)
- ZIDEK, J. V., SHADDICK, G. and TAYLOR, C. G. (2014). Reducing estimation bias in adaptively changing monitoring networks with preferential site selection. *Ann. Appl. Stat.* **8** 1640–1670. [MR3271347](#)
- ZOLTÁN, B.-D., KOVÁCS-LÁNG, E., RÉDEI, T., KERTÉSZ, M. and GARADNAI, J. (2007). Statistical and biological consequences of preferential sampling in phytosociology: Theoretical considerations and a case study. *Folia Geobot.* **42** 141–152.

The Annals of Applied Statistics

Next Issues

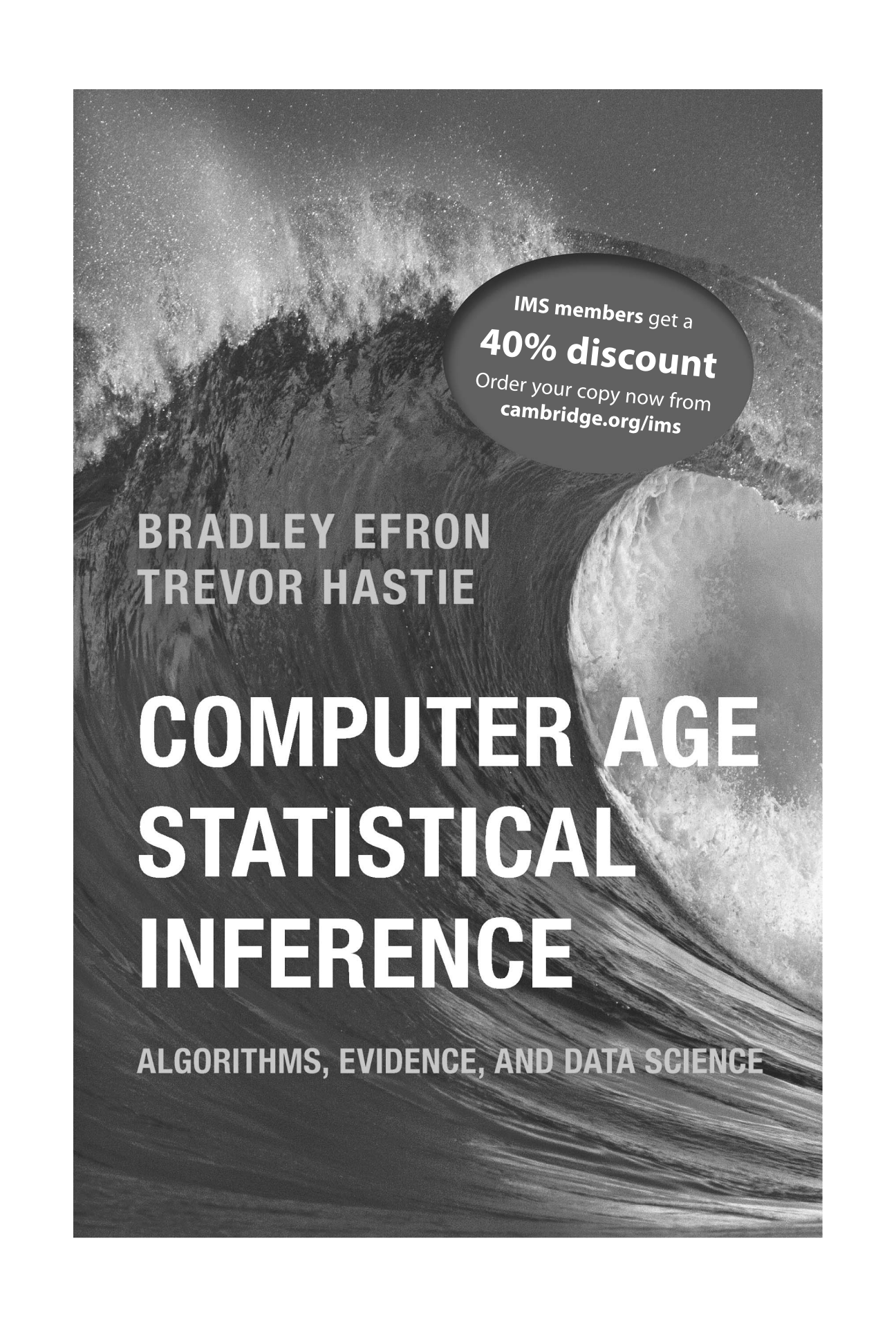
- Bayesian factor models for probabilistic cause of death assessment with verbal autopsies
TSUYOSHI KUNIHAMA, ZEHANG LI, SAMUEL CLARK AND TYLER MCCORMICK
- SHOPPER: A probabilistic model of consumer choice with substitutes and complements
FRANCISCO J. R. RUIZ
- BART with targeted smoothing: An analysis of patient-specific stillbirth risk
JENNIFER E. STARLING, JARED S. MURRAY, CARLOS M. CARVALHO,
RADEK K. BUKOWSKI AND JAMES G. SCOTT
- Efficient real-time monitoring of an emerging influenza pandemic: How feasible?
PAUL JEFFREY BIRRELL, LORENZ WERNISCH, BRIAN D. M. TOM, LEONHARD HELD,
GARETH O. ROBERTS, RICHARD G. PEBODY AND DANIELA DE ANGELIS
- Modeling microbial abundances and dysbiosis with beta-binomial regression
BRYAN D. MARTIN, DANIELA WITTEN AND AMY WILLIS
- Integrative survival analysis with uncertain event times in application to a suicide risk study
WEIJIE WANG, ROBERT ASELTINE, KUN CHEN AND JUN YAN
- A statistical analysis of noisy crowdsourced weather data
ARNAB CHAKRABORTY, SOUMENDRA NATH LAHIRI AND ALYSON WILSON
- The stratified micro-randomized trial design: Sample size considerations for testing nested
causal effects of time-varying treatments WALTER DEMPSEY, PENG LIAO,
SANTOSH KUMAR AND SUSAN MURPHY
- Accounting for uncertainty about past values in probabilistic projections of the total fertility rate
for all countries PEIRAN LIU AND ADRIAN E. RAFTERY
- Bayesian mixed effects models for zero-inflated compositions in microbiome data analysis
BOYU REN, SERGIO BACALLADO, STEFANO FAVARO, TOMMI VATANEN,
CURTIS HUTTENHOWER AND LORENZO TRIPPA
- A hierarchical Bayesian model for predicting ecological interactions using scaled evolutionary
relationships MOHAMAD ELMASRI, MAXWELL JENNER FARRELL,
T. JONATHAN DAVIES AND DAVID ALAN STEPHENS
- Surface temperature monitoring in liver procurement via functional variance change point
analysis ZHENGUO GAO, PANG DU, RAN JIN AND JOHN L. ROBERTSON
- An efficient and computationally robust statistical method for analyzing case-control
mother-offspring pair genetic association studies
HONG ZHANG, BHARAMAR MUKHERJEE, VICTORIA ARTHUR, GANG HU,
HAGIT HOCHNER AND JINBO CHEN
- Regression for copula-linked compound distributions with applications in modeling aggregate
insurance claims PENG SHI AND ZIFENG ZHAO

Continued

The Annals of Applied Statistics

Next Issues–Continued

- Torus graphs for multivariate phase coupling analysis NATALIE KLEIN,
JOSUE ORELLANA, SCOTT BRINCAT, EARL K. MILLER AND ROBERT E. KASS
- Modifying the Chi-square and the CMH test for population genetic inference: Adapting to
over-dispersion KERSTIN SPITZER, MARTA PELIZZOLA AND ANDREAS FUTSCHIK
- TFisher: A powerful truncation and weighting procedure for combining P-values
HONG ZHANG, TIEJUN TONG, JOHN LANDERS AND ZHEYANG WU
- Optimal asset allocation with multivariate Bayesian dynamic linear models
JARED DALE FISHER, DAVIDE PETTENUZZO AND CARLOS M. CARVALHO
- Assessing wage status transition and stagnation using quantile transition regression
CHIH-YUAN HSU, YI-HAU CHEN, RUOH-RONG YU AND TSUNG-WEI HUNG
- A fast particle-based approach for calibrating a 3-d model of the Antarctic ice sheet
BEN SEIYON LEE, MURALI HARAN, ROBERT WILLIAM FULLER,
DAVID POLLARD AND KLAUS KELLER
- Estimating and forecasting the smoking-attributable mortality fraction for both genders jointly
in over 60 countries YICHENG LI AND ADRIAN E. RAFTERY
- Estimating the health effects of environmental mixtures using Bayesian semiparametric
regression and sparsity inducing priors JOSEPH ANTONELLI, MAITREYI MAZUMDAR,
DAVID BELLINGER, DAVID CHRISTIANI, ROBERT WRIGHT AND BRENT COULL



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