

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

- A Hermite–Gaussian based exoplanet radial velocity estimation method
PARKER H. HOLZER, JESSI CISEWSKI-KEHE, DEBRA FISCHER AND LILY ZHAO 527
- Hierarchical integrated spatial process modeling of monotone West Antarctic snow
density curves. . . . PHILIP A. WHITE, DURBAN G. KEELER AND SUMMER RUPPER 556
- Estimating high-resolution Red Sea surface temperature hotspots, using a low-rank
semiparametric spatial model. ARNAB HAZRA AND RAPHAËL HUSER 572
- Learning excursion sets of vector-valued Gaussian random fields for autonomous ocean
sampling. TRYGVE OLAV FOSSUM, CÉDRIC TRAVELLETTI, JO EIDSVIK,
DAVID GINSBOURGER AND KANNA RAJAN 597
- Aggregated pairwise classification of elastic planar shapes
MIN HO CHO, SEBASTIAN KURTEK AND STEVEN N. MACEACHERN 619
- A statistical pipeline for identifying physical features that differentiate classes of 3D
shapes. BRUCE WANG, TIMOTHY SUDIJONO, HENRY KIRVESLAHTI,
TINGRAN GAO, DOUGLAS M. BOYER,
SAYAN MUKHERJEE AND LORIN CRAWFORD 638
- Simultaneous inference of periods and period-luminosity relations for Mira variable stars
SHIYUAN HE, ZHENFENG LIN, WENLONG YUAN,
LUCAS M. MACRI AND JIANHUA Z. HUANG 662
- Scalable penalized spatiotemporal land-use regression for ground-level nitrogen dioxide
KYLE P. MESSIER AND MATTHIAS KATZFUSS 688
- Probabilistic forecasting of the Arctic sea ice edge with contour modeling
HANNAH M. DIRECTOR, ADRIAN E. RAFTERY AND CECILIA M. BITZ 711
- Additive stacking for disaggregate electricity demand forecasting
CHRISTIAN CAPEZZA, BIAGIO PALUMBO, YANNIG GOUDE,
SIMON N. WOOD AND MATTEO FASIOLO 727
- Causal mediation analysis for sparse and irregular longitudinal data. SHUXI ZENG,
STACY ROSENBAUM, SUSAN C. ALBERTS, ELIZABETH A. ARCHIE AND FAN LI 747
- Rapid design of metamaterials via multitarget Bayesian optimization
YANG YANG, CHUNLIN JI AND KE DENG 768
- A continuous-time semi-Markov model for animal movement in a dynamic environment
DEVIN JOHNSON, NOEL PELLAND AND JEREMY STERLING 797
- A Bayesian semiparametric Jolly–Seber model with individual heterogeneity: An
application to migratory mallards at stopover. GUOHUI WU, SCOTT H. HOLAN,
ALEXIS AVRIL AND JONAS WALDENSTRÖM 813
- Inference of large modified Poisson-type graphical models: Application to RNA-seq data
in childhood atopic asthma studies. RONG ZHANG, ZHAO REN,
JUAN C. CELEDÓN AND WEI CHEN 831
- Two-way sparsity for time-varying networks with applications in genomics
THOMAS E. BARTLETT, IOANNIS KOSMIDIS AND RICARDO SILVA 856

continued

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—Continued from front cover

A compositional model to assess expression changes from single-cell RNA-seq data XIUYU MA, KEEGAN KORTHAUER, CHRISTINA KENDZIORSKI AND MICHAEL A. NEWTON	880
Large-scale multiple inference of collective dependence with applications to protein function ROBERT JERNIGAN, KEJUE JIA, ZHAO REN AND WEN ZHOU	902
A Bayesian nonparametric model for inferring subclonal populations from structured DNA sequencing data SHAI HE, AARON SCHEIN, VISHAL SARSANI AND PATRICK FLAHERTY	925
Prediction of the NASH through penalized mixture of logistic regression models MARIE MORVAN, EMILIE DEVIJVER, MADISON GIACOFI AND VALÉRIE MONBET	952
Efficient Bayesian inference of general Gaussian models on large phylogenetic trees PAUL BASTIDE, LAM SI TUNG HO, GUY BAELE, PHILIPPE LEMEY AND MARC A. SUCHARD	971
A covariance-enhanced approach to multitissue joint eQTL mapping with application to transcriptome-wide association studies AARON J. MOLSTAD, WEI SUN AND LI HSU	998
Tensor-variate finite mixture modeling for the analysis of university professor remuneration SHUCHISMITA SARKAR, VOLODYMYR MELNYKOV AND XUWEN ZHU	1017
Inflection points in community-level homeless rates CHRIS GLYNN, THOMAS H. BYRNE AND DENNIS P. CULHANE	1037
Length-biased semicompeting risks models for cross-sectional data: An application to current duration of pregnancy attempt data ALEXANDER C. MCLAIN, SIYUAN GUO, MARIE THOMA AND JIAJIA ZHANG	1054
A bridging model to reconcile statistics based on data from multiple surveys ANDREEA L. ERCIULESCU, JEAN D. OPSOMER AND F. JAY BREIDT	1068

THE ANNALS OF APPLIED STATISTICS

Vol. 15, No. 2, pp. 527–1079 June 2021

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

IMS OFFICERS

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

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

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

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

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

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

IMS PUBLICATIONS

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

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

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

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

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

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

The Annals of Applied Statistics [ISSN 1932-6157 (print); ISSN 1941-7330 (online)], Volume 15, Number 2, June 2021. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage pending at Cleveland, Ohio, and at additional mailing offices.

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

A HERMITE–GAUSSIAN BASED EXOPLANET RADIAL VELOCITY ESTIMATION METHOD

BY PARKER H. HOLZER¹, JESSI CISEWSKI-KEHE², DEBRA FISCHER^{3,*} AND LILY ZHAO^{3,†}

¹Department of Statistics and Data Science, Yale University, parker.holzer@yale.edu

²Department of Statistics, University of Wisconsin–Madison, jjkehe@wisc.edu

³Department of Astronomy, Yale University, * debra.fischer@yale.edu; † lily.zhao@yale.edu

As the first successful technique used to detect exoplanets orbiting distant stars, the radial velocity method aims to detect a periodic Doppler shift in a stellar spectrum due to the star’s motion along the line sight. We introduce a new, mathematically rigorous approach to detect such a signal that accounts for the smooth functional relationship of neighboring wavelengths in the spectrum, minimizes the role of wavelength interpolation, accounts for heteroskedastic noise and easily allows for accurate calculation of the estimated radial velocity standard error. Using Hermite–Gaussian functions, we show that the problem of detecting a Doppler shift in the spectrum can be reduced to linear regression in many settings. A simulation study demonstrates that the proposed method is able to accurately estimate an individual spectrum’s radial velocity with precision below 0.3 m s^{-1} , corresponding to a Doppler shift much smaller than the size of a spectral pixel. Furthermore, the new method outperforms the traditional cross-correlation function approach for estimating the radial velocity by reducing the root mean squared error up to 15 cm s^{-1} . The proposed method is also demonstrated on a new set of observations from the EXtreme PREcision Spectrometer (EXPRES) for the host star 51 Pegasi, and successfully recovers estimates of the planetary companion’s parameters that agree well with previous studies. The method is implemented in the R package *rvmethod*, and supplemental Python code is also available.

REFERENCES

- ANGLADA-ESCUDE, G. and BUTLER, R. P. (2012). The harps-terra project. I. Description of the algorithms, performance, and new measurements on a few remarkable stars observed by harps. *Astrophys. J., Suppl. Ser.* **200** 15.
- ASTUDILLO-DEFRU, N., BONFILS, X., DELFOSSE, X., SÉGRANSAN, D., FORVEILLE, T., BOUCHY, F., GILLON, M., LOVIS, C., MAYOR, M. et al. (2015). The harps search for southern extra-solar planets-xxxvi. Planetary systems and stellar activity of the m dwarfs gj 3293, gj 3341, and gj 3543. *Astron. Astrophys.* **575** A119.
- BARANNE, A., QUELOZ, D., MAYOR, M., ADRIANZYK, G., KNISPEN, G., KOHLER, D., LACROIX, D., MEUNIER, J.-P., RIMBAUD, G. et al. (1996). Elodie: A spectrograph for accurate radial velocity measurements. *Astron. Astrophys. Suppl. Ser.* **119** 373–390.
- BARKER, P. (1984). Ripple correction of high-dispersion iue spectra-blazing echelles. *Astron. J.* **89** 899–903.
- BEDDELL, M., HOGG, D. W., FOREMAN-MACKEY, D., MONTET, B. T. and LUGER, R. (2019). Wobble: A data-driven method for precision radial velocities. *Astron. J.* **158** 164.
- BLACKMAN, R. T., SZYMKOWIAK, A. E., FISCHER, D. A. and JURGENSON, C. A. (2017). Accounting for chromatic atmospheric effects on barycentric corrections. *Astrophys. J.* **837** 18.
- BLACKMAN, R. T., FISCHER, D. A., JURGENSON, C. A., SAWYER, D., MCCracken, T. M., SZYMKOWIAK, A. E., PETERSBURG, R. R., ONG, J. M. J., BREWER, J. M. et al. (2020). Performance verification of the EXtreme PREcision Spectrograph. *Astron. J.* **159** 238.

- BOUCHY, F., PEPE, F. and QUELOZ, D. (2001). Fundamental photon noise limit to radial velocity measurements. *Astron. Astrophys.* **374** 733–739.
- BRANCH, M. A., COLEMAN, T. F. and LI, Y. (1999). A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM J. Sci. Comput.* **21** 1–23. [MR1722103 https://doi.org/10.1137/S1064827595289108](https://doi.org/10.1137/S1064827595289108)
- BUTLER, R. P., MARCY, G. W., WILLIAMS, E., MCCARTHY, C., DOSANJH, P. and VOGT, S. S. (1996). Attaining Doppler precision of 3 m s⁻¹. *Publ. Astron. Soc. Pac.* **108** 500.
- CIURYŁO, R. (1998). Shapes of pressure- and Doppler-broadened spectral lines in the core and near wings. *Phys. Rev. A* **58** 1029.
- DAI, C.-Q., WANG, Y. and LIU, J. (2016). Spatiotemporal Hermite–Gaussian solitons of a (3 + 1)-dimensional partially nonlocal nonlinear Schrödinger equation. *Nonlinear Dynam.* **84** 1157–1161. [MR3486561 https://doi.org/10.1007/s11071-015-2560-9](https://doi.org/10.1007/s11071-015-2560-9)
- DAVIS, A. B., CISEWSKI, J., DUMUSQUE, X., FISCHER, D. A. and FORD, E. B. (2017). Insights on the spectral signatures of stellar activity and planets from pca. *Astrophys. J.* **846** 59.
- DELISLE, J.-B., SÉGRANSAN, D., DUMUSQUE, X., DIAZ, R., BOUCHY, F., LOVIS, C., PEPE, F., UDRY, S., ALONSO, R. et al. (2018). The harps search for southern extra-solar planets-xliii. A compact system of four super-Earth planets orbiting hd 215152. *Astron. Astrophys.* **614** A133.
- DESORT, M., LAGRANGE, A.-M., GALLAND, F., UDRY, S. and MAYOR, M. (2007). Search for exoplanets with the radial-velocity technique: Quantitative diagnostics of stellar activity. *Astron. Astrophys.* **473** 983–993.
- DOPPLER, C. (1842). *Über Das Farbige Licht der Doppelsterne und Einiger Anderer Gestirne des Himmels*. Calve.
- DUMUSQUE, X. (2018). Measuring precise radial velocities on individual spectral lines-I. Validation of the method and application to mitigate stellar activity. *Astron. Astrophys.* **620** A47.
- DUMUSQUE, X., BOISSE, I. and SANTOS, N. (2014). Soap 2.0: A tool to estimate the photometric and radial velocity variations induced by stellar spots and plagues. *Astrophys. J.* **796** 132.
- DUMUSQUE, X., BORSA, F., DAMASSO, M., DÍAZ, R. F., GREGORY, P., HARA, N., HATZES, A., RAJPAUL, V., TUOMI, M. et al. (2017). Radial-velocity fitting challenge-ii. First results of the analysis of the data set. *Astron. Astrophys.* **598** A133.
- EINSTEIN, A. et al. (1905). On the electrodynamics of moving bodies. *Ann. Phys.* **17** 50.
- FISCHER, D. A., ANGLADA-ESCUDE, G., ARRIAGADA, P., BALUEV, R. V., BEAN, J. L., BOUCHY, F., BUCHHAVE, L. A., CARROLL, T., CHAKRABORTY, A. et al. (2016). State of the field: Extreme precision radial velocities. *Publ. Astron. Soc. Pac.* **128** 066001.
- GIGUERE, M. J., FISCHER, D. A., ZHANG, C. X., MATTHEWS, J. M., CAMERON, C. and HENRY, G. W. (2016). A combined spectroscopic and photometric stellar activity study of epsilon eridani. *Astrophys. J.* **824** 150.
- GRAY, D. F. (2005). *The Observation and Analysis of Stellar Photospheres*. Cambridge Univ. Press.
- HALVERSON, S., TERRIEN, R., MAHADEVAN, S., ROY, A., BENDER, C., STEFÁNSSON, G. K., MONSON, A., LEVI, E., HEARTY, F. et al. (2016). A comprehensive radial velocity error budget for next generation Doppler spectrometers. In *Ground-Based and Airborne Instrumentation for Astronomy VI* **9908** page 99086P. International Society for Optics and Photonics.
- HAN, E., WANG, S. X., WRIGHT, J. T., FENG, Y. K., ZHAO, M., FAKHOURI, O., BROWN, J. I. and HANCOCK, C. (2014). Exoplanet orbit database. II. Updates to exoplanets.org. *Publ. Astron. Soc. Pac.* **126** 827.
- HATZES, A. P. (2002). Starspots and exoplanets. *Astron. Nachr.* **323** 392–394.
- HOLZER, P. H. (2021). Source code for “A Hermite–Gaussian based exoplanet radial velocity estimation method.” <https://doi.org/10.1214/20-AOAS1406SUPPB>.
- HOLZER, P. H., CISEWSKI-KEHE, J., FISCHER, D. and ZHAO, L. (2021). Supplement to “A Hermite–Gaussian based exoplanet radial velocity estimation method.” <https://doi.org/10.1214/20-AOAS1406SUPPA>.
- ISAACSON, H. and FISCHER, D. (2010). Chromospheric activity and jitter measurements for 2630 stars on the California planet search. *Astrophys. J.* **725** 875.
- JOHNSTON, W. (2014). The weighted Hermite polynomials form a basis for $L^2(\mathbb{R})$. *Amer. Math. Monthly* **121** 249–253. [MR3168998 https://doi.org/10.4169/amer.math.monthly.121.03.249](https://doi.org/10.4169/amer.math.monthly.121.03.249)
- JURGENSON, C., FISCHER, D., MCCracken, T., SAWYER, D., SZYMOWIAK, A., DAVIS, A., MULLER, G. and SANTORO, F. (2016). Expres: A next generation rv spectrograph in the search for Earth-like worlds. In *Ground-Based and Airborne Instrumentation for Astronomy VI* **9908** page 99086T. International Society for Optics and Photonics.
- LAGRANGE, A.-M., DESORT, M. and MEUNIER, N. (2010). Using the sun to estimate Earth-like planets detection capabilities-I. Impact of cold spots. *Astron. Astrophys.* **512** A38.
- LANCZOS, C. (1938). Trigonometric interpolation of empirical and analytical functions. *J. Math. Phys.* **17** 123–199.

- LEET, C., FISCHER, D. A. and VALENTI, J. A. (2019). Towards a self-calibrating, empirical, light-weight model for tellurics in high-resolution spectra. *Astron. J.* **157** 187.
- MARCY, G. W., BUTLER, R. P., WILLIAMS, E., BILDSTEN, L., GRAHAM, J. R., GHEZ, A. M. and JERNIGAN, J. G. (1997). The planet around 51 pegasi. *Astrophys. J.* **481** 926.
- MARHIC, M. E. (1978). Oscillating Hermite–Gaussian wave functions of the harmonic oscillator. *Lett. Nuovo Cimento* (2) **22** 376–378. [MR0495931](#)
- MAYOR, M. and QUELOZ, D. (1995). A Jupiter-mass companion to a solar-type star. *Nature* **378** 355.
- MÉSZÁROS, S. and PRIETO, C. A. (2013). On the interpolation of model atmospheres and high-resolution synthetic stellar spectra. *Mon. Not. R. Astron. Soc.* **430** 3285–3291.
- MEUNIER, N., DESORT, M. and LAGRANGE, A.-M. (2010). Using the sun to estimate Earth-like planets detection capabilities-ii. Impact of plagues. *Astron. Astrophys.* **512** A39.
- MORÉ, J. J. (1978). The Levenberg–Marquardt algorithm: Implementation and theory. In *Numerical Analysis (Proc. 7th Biennial Conf., Univ. Dundee, Dundee, 1977). Lecture Notes in Math.* **630** 105–116. [MR0483445](#)
- PAULSON, D. B., COCHRAN, W. D. and HATZES, A. P. (2004). Searching for planets in the hyades. V. Limits on planet detection in the presence of stellar activity. *Astron. J.* **127** 3579.
- PEPE, F., MAYOR, M., GALLAND, F., NAEF, D., QUELOZ, D., SANTOS, N., UDRY, S. and BURNET, M. (2002). The coralie survey for southern extra-solar planets vii-two short-period saturnian companions to hd 108147 and hd 168746. *Astron. Astrophys.* **388** 632–638.
- PETERSBURG, R. R., ONG, J. M. J., ZHAO, L. L., BLACKMAN, R. T., BREWER, J. M., BUCHHAVE, L. A., CABOT, S. H. C., DAVIS, A. B., JURGENSON, C. A. et al. (2020). An extreme precision radial velocity pipeline: First radial velocities from EXPRES. *Astron. J.* **159** 187.
- PLANCK, M. (1901). On the law of distribution of energy in the normal spectrum. *Ann. Phys.* **4** 1.
- QUELOZ, D., HENRY, G., SIVAN, J., BALIUNAS, S., BEUZIT, J., DONAHUE, R., MAYOR, M., NAEF, D., PERRIER, C. et al. (2001). No planet for hd 166435. *Astron. Astrophys.* **379** 279–287.
- RAJPAUL, V. M., AIGRAIN, S. and BUCHHAVE, L. A. (2020). A robust, template-free approach to precise radial velocity extraction. *Mon. Not. R. Astron. Soc.* **492** 3960–3983.
- RAJPAUL, V., AIGRAIN, S., OSBORNE, M. A., REECE, S. and ROBERTS, S. (2015). A Gaussian process framework for modelling stellar activity signals in radial velocity data. *Mon. Not. R. Astron. Soc.* **452** 2269–2291.
- RIFFEL, R. A. (2010). Profit: A new alternative for emission-line profile fitting. *Astrophys. Space Sci.* **327** 239–244.
- RIMMELE, T. R. and RADICK, R. R. (1998). Solar adaptive optics at the national solar observatory. In *Adaptive Optical System Technologies* **3353** 72–82. International Society for Optics and Photonics.
- SAAR, S. H. and DONAHUE, R. A. (1997). Activity-related radial velocity variation in cool stars. *Astrophys. J.* **485** 319.
- SIMOLA, U., DUMUSQUE, X. and CISEWSKI-KEHE, J. (2019). Measuring precise radial velocities and cross-correlation function line-profile variations using a skew normal density. *Astron. Astrophys.* **622** A131.
- SNEDEN, C., BEAN, J., IVANS, I., LUCATELLO, S. and SOBECK, J. (2012). Moog: Lte line analysis and spectrum synthesis. Astrophysics Source Code Library.
- TUOMI, M., ANGLADA-ESCUDE, G., GERLACH, E., JONES, H. R., REINERS, A., RIVERA, E. J., VOGT, S. S. and BUTLER, R. P. (2013). Habitable-zone super-Earth candidate in a six-planet system around the k2. 5v star hd 40307. *Astron. Astrophys.* **549** A48.
- WANG, J. and FORD, E. B. (2011). On the eccentricity distribution of short-period single-planet systems. *Mon. Not. R. Astron. Soc.* **418** 1822–1833.
- WRIGHT, J. and EASTMAN, J. (2014). Barycentric corrections at 1 cm s⁻¹ for precise Doppler velocities. *Publ. Astron. Soc. Pac.* **126** 838.
- XU, X., CISEWSKI-KEHE, J., DAVIS, A. B., FISCHER, D. A. and BREWER, J. M. (2019). Modeling the echelle spectra continuum with alpha shapes and local regression fitting. *Astron. J.* **157** 243.

HIERARCHICAL INTEGRATED SPATIAL PROCESS MODELING OF MONOTONE WEST ANTARCTIC SNOW DENSITY CURVES

BY PHILIP A. WHITE¹, DURBAN G. KEELER^{2,*} AND SUMMER RUPPER^{2,†}

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

²*Department of Geography, University of Utah, *durban.keeler@gmail.com; †summer.rupper@geog.utah.edu*

Snow density estimates below the surface, used with airplane-acquired ice-penetrating radar measurements, give a site-specific history of snow water accumulation. Because it is infeasible to drill snow cores across all of Antarctica to measure snow density and because it is critical to understand how climatic changes are affecting the world's largest freshwater reservoir, we develop methods that enable snow density estimation with uncertainty in regions where snow cores have not been drilled.

In inland West Antarctica, snow density increases monotonically as a function of depth, except for possible microscale variability or measurement error, and it cannot exceed the density of ice. We present a novel class of integrated spatial process models that allow interpolation of monotone snow density curves. For computational feasibility we construct the space-depth process through kernel convolutions of log-Gaussian spatial processes. We discuss model comparison, model fitting and prediction. Using this model, we extend estimates of snow density beyond the depth of the original core and estimate snow density curves where snow cores have not been drilled. Along flight lines with ice-penetrating radar, we use interpolated snow density curves to estimate recent water accumulation and find predominantly decreasing water accumulation over recent decades.

REFERENCES

- ALBERT, M., SHUMAN, C., COURVILLE, Z., BAUER, R., FAHNESTOCK, M. and SCAMBOS, T. (2004). Extreme firm metamorphism: Impact of decades of vapor transport on near-surface firm at a low-accumulation glazed site on the East Antarctic plateau. *Annals of Glaciology* **39** 73–78.
- ALLEY, R. B., SALTZMAN, E. S., CUFFEY, K. M. and FITZPATRICK, J. J. (1990). Summertime formation of Depth Hoar in central Greenland. *Geophys. Res. Lett.* **17** 2393–2396.
- BANTA, J. R., MCCONNELL, J. R., FREY, M. M., BALES, R. C. and TAYLOR, K. (2008). Spatial and temporal variability in snow accumulation at the West Antarctic ice sheet divide over recent centuries. *J. Geophys. Res., Atmos.* **113** D23102.
- BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions*. Wiley, New York.
- BROMWICH, D. H., NICOLAS, J. P., MONAGHAN, A. J., LAZZARA, M. A., KELLER, L. M., WEIDNER, G. A. and WILSON, A. B. (2013). Central West Antarctica among the most rapidly warming regions on Earth. *Nat. Geosci.* **6** 139–145.
- BROWN, T. A. (1974). Admissible scoring systems for continuous distributions. Technical Report No. P-5235, The Rand Corporation, Santa Monica, California.
- BURGENER, L., RUPPER, S., KOENIG, L., FORSTER, R., CHRISTENSEN, W. F., WILLIAMS, J., KOUTNIK, M., MIEGE, C., STEIG, E. J. et al. (2013). An observed negative trend in West Antarctic accumulation rates from 1975 to 2010: Evidence from new observed and simulated records. *J. Geophys. Res., Atmos.* **118** 4205–4216.
- CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. et al. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76**.
- CUFFEY, K. M. and PATERSON, W. S. B. (2010). *The Physics of Glaciers*. Academic Press, San Diego.
- EISEN, O., FREZZOTTI, M., GENTHON, C., ISAKSSON, E., MAGAND, O., VAN DEN BROEKE, M. R., DIXON, D. A., EKAYKIN, A., HOLMLUND, P. et al. (2008). Ground-based measurements of spatial and temporal variability of snow accumulation in East Antarctica. *Reviews of Geophysics* **46** RG2001.

- FRYER, M. (1976). Some errors associated with the non-parametric estimation of density functions. *IMA J. Appl. Math.* **18** 371–380.
- FUJITA, S., MAENO, H., URATSUKA, S., FURUKAWA, T., MAE, S., FUJII, Y. and WATANABE, O. (1999). Nature of radio echo layering in the Antarctic ice sheet detected by a two-frequency experiment. *J. Geophys. Res., Solid Earth* **104** 13013–13024.
- GELFAND, A. E. and KUO, L. (1991). Nonparametric Bayesian bioassay including ordered polytomous response. *Biometrika* **78** 657–666. [MR1130934 https://doi.org/10.1093/biomet/78.3.657](https://doi.org/10.1093/biomet/78.3.657)
- GELFAND, A. E., KIM, H.-J., SIRMANS, C. F. and BANERJEE, S. (2003). Spatial modeling with spatially varying coefficient processes. *J. Amer. Statist. Assoc.* **98** 387–396. [MR1995715 https://doi.org/10.1198/0162145030000170](https://doi.org/10.1198/0162145030000170)
- GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. [MR2345548 https://doi.org/10.1198/016214506000001437](https://doi.org/10.1198/016214506000001437)
- 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. <https://doi.org/10.1002/ecy.1674>
- HERRON, M. M. and LANGWAY, C. C. (1980). Firn densification: An empirical model. *Journal of Glaciology* **25** 373–385.
- HIGDON, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environ. Ecol. Stat.* **5** 173–190.
- HIGDON, D. (2002). Space and space–time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* 37–56. Springer, London. [MR2059819](https://doi.org/10.1007/978-1-4020-0819-9_3)
- HÖRHHOLD, M. W., KIPFSTUHL, S., WILHELMS, F., FREITAG, J. and FRENZEL, A. (2011). The densification of layered polar firn. *J. Geophys. Res., Earth Surf.* **116**.
- IPCC (2013). Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge Univ. Press, Cambridge, United Kingdom and New York, NY, USA.
- KEELER, D. G., RUPPER, S. B., FORSTER, R. and MIÈGE, C. (2020). A probabilistic automated isochrone picking routine to derive annual surface mass balance from radar echograms. *IEEE Trans. Geosci. Remote Sens.*
- KOENIG, L., FORSTER, R., BRUCKER, L. and MILLER, J. (2014). Remote sensing of accumulation over the Greenland and Antarctic ice sheets. In *Remote Sensing of the Cryosphere* 157–186. Wiley-Blackwell, New York.
- KRÜGER, F., LERCH, S., THORARINSDOTTIR, T. L. and GNEITING, T. (2020). Predictive inference based on Markov chain Monte Carlo output. *Int. Stat. Rev.*
- LAMOREY, G. and COOPER, T. (2002). Siple dome WAISCORES density data for shallow cores.
- LANG, S. and BREZGER, A. (2004). Bayesian P-splines. *J. Comput. Graph. Statist.* **13** 183–212. [MR2044877 https://doi.org/10.1198/1061860043010](https://doi.org/10.1198/1061860043010)
- LENAERTS, J. T. M., VAN DEN BROEKE, M. R., VAN DE BERG, W. J., VAN MEIJGAARD, E. and MUNNEKE, P. K. (2012). A new, high-resolution surface mass balance map of Antarctica (1979–2010) based on regional atmospheric climate modeling. *Geophys. Res. Lett.* **39** L04501.
- LI, J. and ZWALLY, H. J. (2011). Modeling of firn compaction for estimating ice-sheet mass change from observed ice-sheet elevation change. *Annals of Glaciology* **52** 1–7.
- LIN, L. and DUNSON, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika* **101** 303–317. [MR3215349 https://doi.org/10.1093/biomet/ast063](https://doi.org/10.1093/biomet/ast063)
- MARRON, J. S. and WAND, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** 712–736. [MR1165589 https://doi.org/10.1214/aos/1176348653](https://doi.org/10.1214/aos/1176348653)
- MATHESON, J. E. and WINKLER, R. L. (1976). Scoring rules for continuous probability distributions. *Manage. Sci.* **22** 1087–1096.
- MAYEWSKI, P. A., FREZZOTTI, M., BERTLER, N., OMMEN, T. V., HAMILTON, G., JACKA, T. H., WELCH, B., FREY, M., DAHE, Q. et al. (2005). The international trans-Antarctic scientific expedition (ITASE): An overview. *Annals of Glaciology* **41** 180–185.
- MEDLEY, B., JOUGHIN, I., SMITH, B. E., DAS, S. B., STEIG, E. J., CONWAY, H., GOGINENI, S., LEWIS, C., CRISCITIELLO, A. S. et al. (2014). Constraining the recent mass balance of pine island and thwaites glaciers, West Antarctica, with airborne observations of snow accumulation. *Cryosphere* **8** 1375–1392.
- MEYER, M. C. (2008). Inference using shape-restricted regression splines. *Ann. Appl. Stat.* **2** 1013–1033. [MR2516802 https://doi.org/10.1214/08-AOAS167](https://doi.org/10.1214/08-AOAS167)
- MIÈGE, C., FORSTER, R. R., BOX, J. E., BURGESS, E. W., MCCONNELL, J. R., PASTERIS, D. R. and SPIKES, V. B. (2013). Southeast Greenland high accumulation rates derived from firn cores and ground-penetrating radar. *Annals of Glaciology* **54** 322–332.

- MÜLLER, H.-G. and SCHMITT, T. (1988). Kernel and probit estimates in quantal bioassay. *J. Amer. Statist. Assoc.* **83** 750–759. [MR0963803](#)
- NEELON, B. and DUNSON, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics* **60** 398–406. [MR2066274](#) <https://doi.org/10.1111/j.0006-341X.2004.00184.x>
- PADEN, J., LI, J., LEUSCHEN, C., RODRIGUEZ-MORALES, F. and HALE, R. (2019). IceBridge snow radar L1B geolocated radar echo strength profiles, Version 2.
- RAMSAY, J. O. (1988). Monotone regression splines in action. *Statist. Sci.* **3** 425–441.
- RIIHIMÄKI, J. and VEHTARI, A. (2010). Gaussian processes with monotonicity information. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 645–652.
- ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Wiley, Chichester. [MR0961262](#)
- TANNER, M. A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd ed. *Springer Series in Statistics*. Springer, New York. [MR1396311](#) <https://doi.org/10.1007/978-1-4612-4024-2>
- TEDESCO, M., PARRINELLO, T., WEBB, C. and MARKUS, T. (2014). Remote sensing missions and the cryosphere. In *Remote Sensing of the Cryosphere* 382–392. Wiley, New York.
- VERJANS, V., LEESON, A. A., NEMETH, C., STEVENS, C. M., KUIPERS MUNNEKE, P., NOËL, B. and WESSEM, J. M. V. (2020). Bayesian calibration of firn densification models. *Cryosphere Discuss.* 1–23.
- WHITE, P., KEELER, D. and RUPPER, S. (2021). Supplement to “Hierarchical integrated spatial process modeling of monotone West Antarctic snow density curves.” <https://doi.org/10.1214/21-AOAS1443SUPPA>, <https://doi.org/10.1214/21-AOAS1443SUPPB>
- WHITE, P. A., REESE, C. S., CHRISTENSEN, W. F. and RUPPER, S. (2019). A model for Antarctic surface mass balance and ice core site selection. *Environmetrics* **30** e2579, 19. [MR4039225](#) <https://doi.org/10.1002/env.2579>
- WHITE, P., KEELER, D., SHEANSHANG, D. and RUPPER, S. (2020). Improving interpretable piecewise linear models through hierarchical spatial and functional smoothing. Preprint. Available at [arXiv:2006.09329](#).

ESTIMATING HIGH-RESOLUTION RED SEA SURFACE TEMPERATURE HOTSPOTS, USING A LOW-RANK SEMIPARAMETRIC SPATIAL MODEL

BY ARNAB HAZRA^{*} AND RAPHAËL HUSER[†]

Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), ^{}arnab.hazra@kaust.edu.sa; [†]raphael.huser@kaust.edu.sa*

In this work, we estimate extreme sea surface temperature (SST) hotspots, that is, high threshold exceedance regions, for the Red Sea, a vital region of high biodiversity. We analyze high-resolution satellite-derived SST data comprising daily measurements at 16,703 grid cells across the Red Sea over the period 1985–2015. We propose a semiparametric Bayesian spatial mixed-effects linear model with a flexible mean structure to capture spatially-varying trend and seasonality, while the residual spatial variability is modeled through a Dirichlet process mixture (DPM) of low-rank spatial Student's t processes (LTPs). By specifying cluster-specific parameters for each LTP mixture component, the bulk of the SST residuals influence tail inference and hotspot estimation only moderately. Our proposed model has a nonstationary mean, covariance, and tail dependence, and posterior inference can be drawn efficiently through Gibbs sampling. In our application, we show that the proposed method outperforms some natural parametric and semiparametric alternatives. Moreover, we show how hotspots can be identified, and we estimate extreme SST hotspots for the whole Red Sea, projected until the year 2100, based on the Representative Concentration Pathways 4.5 and 8.5. The estimated 95% credible region, for joint high threshold exceedances include large areas covering major endangered coral reefs in the southern Red Sea.

REFERENCES

- ALEXANDERIAN, A. (2015). A brief note on the Karhunen–Loève expansion. Preprint. Available at [arXiv:1509.07526](https://arxiv.org/abs/1509.07526).
- ANDERES, E., HUSER, R., NYCHKA, D. and CORAM, M. (2013). Nonstationary positive definite tapering on the plane. *J. Comput. Graph. Statist.* **22** 848–865. [MR3173746 https://doi.org/10.1080/10618600.2012.729982](https://doi.org/10.1080/10618600.2012.729982)
- BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 825–848. [MR2523906 https://doi.org/10.1111/j.1467-9868.2008.00663.x](https://doi.org/10.1111/j.1467-9868.2008.00663.x)
- BEAKY, M. M., SCHERRER, R. J. and VILLUMSEN, J. V. (1992). Topology of large-scale structure in seeded hot dark matter models. *Astrophys. J.* **387** 443–448.
- BERUMEN, M. L., HOEY, A. S., BASS, W. H., BOUWMEESTER, J., CATANIA, D., COCHRAN, J. E., KHALIL, M. T., MIYAKE, S., MUGHAL, M. et al. (2013). The status of coral reef ecology research in the Red Sea. *Coral Reefs* **32** 737–748.
- BOLIN, D. and LINDGREN, F. (2015). Excursion and contour uncertainty regions for latent Gaussian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 85–106. [MR3299400 https://doi.org/10.1111/rssb.12055](https://doi.org/10.1111/rssb.12055)
- BOPP, G., SHABY, B. A. and HUSER, R. (2021). A hierarchical max-infinitely divisible spatial model for extreme precipitation. *J. Amer. Statist. Assoc.* **116** 93–106. <https://doi.org/10.1080/01621459.2020.1750414>
- BOPP, G. P., SHABY, B. A., FOREST, C. E. and MEJÍA, A. (2020). Projecting flood-inducing precipitation with a Bayesian analogue model. *J. Agric. Biol. Environ. Stat.* **25** 229–249. [MR4089687 https://doi.org/10.1007/s13253-020-00391-6](https://doi.org/10.1007/s13253-020-00391-6)
- CASTRO-CAMILO, D., HUSER, R. and RUE, H. (2019). A spliced gamma-generalized Pareto model for short-term extreme wind speed probabilistic forecasting. *J. Agric. Biol. Environ. Stat.* **24** 517–534. [MR3996457 https://doi.org/10.1007/s13253-019-00369-z](https://doi.org/10.1007/s13253-019-00369-z)

Key words and phrases. Bayesian inference, Dirichlet process mixture model, extreme event, low-rank method, Nonstationary mean, covariance and tail dependence, sea surface temperature data, Student's t process.

- CASTRUCCIO, S., HUSER, R. and GENTON, M. G. (2016). High-order composite likelihood inference for max-stable distributions and processes. *J. Comput. Graph. Statist.* **25** 1212–1229. [MR3572037](#) <https://doi.org/10.1080/10618600.2015.1086656>
- CHENG, L., AGHAKOUCHAK, A., GILLELAND, E. and KATZ, R. W. (2014). Non-stationary extreme value analysis in a changing climate. *Clim. Change* **127** 353–369.
- COOLEY, D. and SAIN, S. R. (2010). Spatial hierarchical modeling of precipitation extremes from a regional climate model. *J. Agric. Biol. Environ. Stat.* **15** 381–402. [MR2787265](#) <https://doi.org/10.1007/s13253-010-0023-9>
- CRAIGMILE, P. F., CRESSIE, N., SANTNER, T. J. and RAO, Y. (2005). A loss function approach to identifying environmental exceedances. *Extremes* **8** 143–159. [MR2275915](#) <https://doi.org/10.1007/s10687-006-7964-y>
- CRESSIE, N. A. C., ZHANG, J. and CRAIGMILE, P. F. (2005). Geostatistical prediction of spatial extremes and their extent. In *Geostatistics for Environmental Applications, Proceedings of the Fifth Conference on Geostatistics for Environmental Applications* (P. Renard, H. Demougeot-Renard and R. Froidevaux, eds.) 27–37. Springer, Berlin.
- DAVISON, A. C. and HUSER, R. (2015). Statistics of extremes. *Annu. Rev. Stat. Appl.* **2** 203–235. <https://doi.org/10.1146/annurev-statistics-010814-020133>
- DAVISON, A. C., HUSER, R. and THIBAUD, E. (2013). Geostatistics of dependent and asymptotically independent extremes. *Math. Geosci.* **45** 511–529. [MR3079649](#) <https://doi.org/10.1007/s11004-013-9469-y>
- DAVISON, A., HUSER, R. and THIBAUD, E. (2019). Spatial extremes. In *Handbook of Environmental and Ecological Statistics M. Fuentes, J. A. Hoeting and R. L. Smith, eds.* Chapman & Hall/CRC Handb. Mod. Stat. Methods 711–744. CRC Press, Boca Raton, FL. [MR3889918](#)
- DAVISON, A. C., PADOAN, S. A. and RIBATET, M. (2012). Statistical modeling of spatial extremes. *Statist. Sci.* **27** 161–186. [MR2963980](#) <https://doi.org/10.1214/11-STS376>
- DE FONDEVILLE, R. and DAVISON, A. C. (2018). High-dimensional peaks-over-threshold inference. *Biometrika* **105** 575–592. [MR3842886](#) <https://doi.org/10.1093/biomet/asy026>
- DEMARTA, S. and MCNEIL, A. J. (2005). The t copula and related copulas. *Int. Stat. Rev.* **73** 111–129.
- DONLON, C. J., MARTIN, M., STARK, J., ROBERTS-JONES, J., FIEDLER, E. and WIMMER, W. (2012). The operational sea surface temperature and sea ice analysis (OSTIA) system. *Remote Sens. Environ.* **116** 140–158.
- DUAN, J. A., GUINDANI, M. and GELFAND, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika* **94** 809–825. [MR2416794](#) <https://doi.org/10.1093/biomet/asm071>
- DUTTA, S. and BHATTACHARYA, S. (2014). Markov chain Monte Carlo based on deterministic transformations. *Stat. Methodol.* **16** 100–116. [MR3110892](#) <https://doi.org/10.1016/j.stamet.2013.08.006>
- EKLUNDH, L. and OLSSON, L. (2003). Vegetation index trends for the African Sahel 1982–1999. *Geophys. Res. Lett.* **30** 1430.
- FIX, M. J., COOLEY, D., SAIN, S. R. and TEBALDI, C. (2018). A comparison of U.S. precipitation extremes under RCP8.5 and RCP4.5 with an application of pattern scaling. *Clim. Change* **146** 335–347.
- FRENCH, J. P. and HOETING, J. A. (2016). Credible regions for exceedance sets of geostatistical data. *Environmetrics* **27** 4–14. [MR3457233](#) <https://doi.org/10.1002/env.2371>
- FRENCH, J. P. and SAIN, S. R. (2013). Spatio-temporal exceedance locations and confidence regions. *Ann. Appl. Stat.* **7** 1421–1449. [MR3127953](#) <https://doi.org/10.1214/13-AOAS631>
- FUENTES, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *J. Amer. Statist. Assoc.* **102** 321–331. [MR2345545](#) <https://doi.org/10.1198/016214506000000852>
- FURRER, R., GENTON, M. G. and NYCHKA, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15** 502–523. [MR2291261](#) <https://doi.org/10.1198/106186006X132178>
- FURRER, R., KNUTTI, R., SAIN, S., NYCHKA, D. and MEEHL, G. (2007). Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophys. Res. Lett.* **34** L06711.
- GELFAND, A. E., KOTTAS, A. and MACEACHERN, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100** 1021–1035. [MR2201028](#) <https://doi.org/10.1198/016214504000002078>
- GELFAND, A. E. and SCHLIEP, E. M. (2016). Spatial statistics and Gaussian processes: A beautiful marriage. *Spat. Stat.* **18** 86–104. [MR3573271](#) <https://doi.org/10.1016/j.spasta.2016.03.006>
- GENEVIER, L. G. C., JAMIL, T., RAITOS, D. E., KROKOS, G. and HOTEIT, I. (2019). Marine heatwaves reveal coral reef zones susceptible to bleaching in the Red Sea. *Glob. Change Biol.* **25** 2338–2351. <https://doi.org/10.1111/gcb.14652>
- GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference. Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. [MR3587782](#) <https://doi.org/10.1017/9781139029834>
- GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. [MR2345548](#) <https://doi.org/10.1198/016214506000001437>

- GNEITING, T. and RANJAN, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econom. Statist.* **29** 411–422. [MR2848512](#) <https://doi.org/10.1198/jbes.2010.08110>
- HAZRA, A. and HUSER, R. (2021a). Supplement to “Estimating high-resolution Red Sea surface temperature hotspots, using a low-rank semiparametric spatial model.” <https://doi.org/10.1214/20-AOAS1418SUPPA>.
- HAZRA, A. and HUSER, R. (2021b). R code for “Estimating high-resolution Red Sea surface temperature hotspots, using a low-rank semiparametric spatial model.” <https://doi.org/10.1214/20-AOAS1418SUPPB>.
- HAZRA, A., REICH, B. J. and STAICU, A.-M. (2020). A multivariate spatial skew- t process for joint modeling of extreme precipitation indexes. *Environmetrics* **31** e2602, 19. [MR4098481](#) <https://doi.org/10.1002/env.2602>
- HAZRA, A., REICH, B. J., SHABY, B. A. and STAICU, A.-M. (2018). A semiparametric Bayesian model for spatiotemporal extremes. Preprint. Available at [arXiv:1812.11699](https://arxiv.org/abs/1812.11699).
- 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 analysing large spatial data. *J. Agric. Biol. Environ. Stat.* **24** 398–425.
- HIGDON, D. (2002). Space and space–time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* 37–56. Springer, London. [MR2059819](#)
- HUSER, R. (2021). Editorial: EVA 2019 data competition on spatio-temporal prediction of Red Sea surface temperature extremes. *Extremes* **24** 91–104. <https://doi.org/10.1007/s10687-019-00369-9>
- HUSER, R., OPITZ, T. and THIBAUD, E. (2017). Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures. *Spat. Stat.* **21** 166–186. [MR3692183](#) <https://doi.org/10.1016/j.spasta.2017.06.004>
- HUSER, R. and WADSWORTH, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *J. Amer. Statist. Assoc.* **114** 434–444. [MR3941266](#) <https://doi.org/10.1080/01621459.2017.1411813>
- HUSER, R. and WADSWORTH, J. L. (2020). Advances in statistical modeling of spatial extremes. *Wiley Interdiscip. Rev.: Comput. Stat.* e1537. <https://doi.org/10.1002/wics.1537>
- HUSER, R., DOMBRY, C., RIBATET, M. and GENTON, M. G. (2019). Full likelihood inference for max-stable data. *Stat* **8** e218, 14. [MR3938281](#) <https://doi.org/10.1002/sta4.218>
- LEWANDOWSKA, A. M., BOYCE, D. G., HOFMANN, M., MATTHIESSEN, B., SOMMER, U. and WORM, B. (2014). Effects of sea surface warming on marine plankton. *Ecol. Lett.* **17** 614–623.
- MEJIA, A. F., YUE, Y., BOLIN, D., LINDGREN, F. and LINDQUIST, M. A. (2020). A Bayesian general linear modeling approach to Cortical surface fMRI data analysis. *J. Amer. Statist. Assoc.* **115** 501–520. [MR4107654](#) <https://doi.org/10.1080/01621459.2019.1611582>
- MORRIS, S. A., REICH, B. J., THIBAUD, E. and COOLEY, D. (2017). A space–time skew- t model for threshold exceedances. *Biometrics* **73** 749–758. [MR3713109](#) <https://doi.org/10.1111/biom.12644>
- OPITZ, T. (2013). Extremal t processes: Elliptical domain of attraction and a spectral representation. *J. Multivariate Anal.* **122** 409–413. [MR3189331](#) <https://doi.org/10.1016/j.jmva.2013.08.008>
- OPITZ, T., HUSER, R., BAKKA, H. and RUE, H. (2018). INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. *Extremes* **21** 441–462. [MR3855716](#) <https://doi.org/10.1007/s10687-018-0324-x>
- PADOAN, S. A., RIBATET, M. and SISSON, S. A. (2010). Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc.* **105** 263–277. [MR2757202](#) <https://doi.org/10.1198/jasa.2009.tm08577>
- PAPADAKIS, M., TSAGRIS, M., DIMITRIADIS, M., TSAMARDINOS, I., FASIOLO, M., BOR-BOUDAKIS, G. and BURKARDT, J. (2017). Rfast: Fast R functions. R Package Version 1.
- QIU, Y. and MEI, J. (2016). Package ‘rARPACK’.
- REASER, J. K., POMERANCE, R. and THOMAS, P. O. (2000). Coral bleaching and global climate change: Scientific findings and policy recommendations. *Conserv. Biol.* **14** 1500–1511.
- REICH, B. J. and SHABY, B. A. (2012). A hierarchical max-stable spatial model for extreme precipitation. *Ann. Appl. Stat.* **6** 1430–1451. [MR3058670](#) <https://doi.org/10.1214/12-AOAS591>
- RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Monographs on Statistics and Applied Probability **104**. CRC Press/CRC, Boca Raton, FL. [MR2130347](#) <https://doi.org/10.1201/9780203492024>
- RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. [MR2649602](#) <https://doi.org/10.1111/j.1467-9868.2008.00700.x>
- SANG, H. and GELFAND, A. E. (2009). Hierarchical modeling for extreme values observed over space and time. *Environ. Ecol. Stat.* **16** 407–426. [MR2749848](#) <https://doi.org/10.1007/s10651-007-0078-0>
- SANG, H. and GELFAND, A. E. (2010). Continuous spatial process models for spatial extreme values. *J. Agric. Biol. Environ. Stat.* **15** 49–65. [MR2755384](#) <https://doi.org/10.1007/s13253-009-0010-1>
- SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. [MR1309433](#)
- SHALTOUT, M. (2019). Recent sea surface temperature trends and future scenarios for the Red Sea. *Oceanologia* **61** 484–504.

- SMITH, R. L. (1990). Max-stable processes and spatial extremes. Unpublished manuscript.
- STEIN, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer, New York. [MR1697409](#) <https://doi.org/10.1007/978-1-4612-1494-6>
- STEIN, M. L., CHI, Z. and WELTY, L. J. (2004). Approximating likelihoods for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 275–296. [MR2062376](#) <https://doi.org/10.1046/j.1369-7412.2003.05512.x>
- THIBAUD, E. and OPITZ, T. (2015). Efficient inference and simulation for elliptical Pareto processes. *Biometrika* **102** 855–870. [MR3431558](#) <https://doi.org/10.1093/biomet/asv045>
- VECCHIA, A. V. (1988). Estimation and model identification for continuous spatial processes. *J. Roy. Statist. Soc. Ser. B* **50** 297–312. [MR0964183](#)
- WIKLE, C. K. (2010). Low-rank representations for spatial processes. In *Handbook of Spatial Statistics* (A. E. Gelfand, P. Diggle, P. Guttorp and M. Fuentes, eds.). Chapman & Hall/CRC Handb. Mod. Stat. Methods 107–118. CRC Press, Boca Raton, FL. [MR2730946](#) <https://doi.org/10.1201/9781420072884-c8>
- WIKLE, C. K. and CRESSIE, N. (1999). A dimension-reduced approach to space–time Kalman filtering. *Biometrika* **86** 815–829. [MR1741979](#) <https://doi.org/10.1093/biomet/86.4.815>
- WU, Y. and GHOSAL, S. (2010). The L_1 -consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *J. Multivariate Anal.* **101** 2411–2419. [MR2719871](#) <https://doi.org/10.1016/j.jmva.2010.06.012>

LEARNING EXCURSION SETS OF VECTOR-VALUED GAUSSIAN RANDOM FIELDS FOR AUTONOMOUS OCEAN SAMPLING

BY TRYGVE OLAV FOSSUM¹, CÉDRIC TRAVELLETI^{2,*}, JO EIDSVIK³,
DAVID GINSBOURGER^{2,†} AND KANNA RAJAN⁴

¹Department of Marine Technology, Norwegian University of Science and Technology (NTNU), trygve.o.fossum@ntnu.no

²Institute of Mathematical Statistics and Actuarial Science, University of Bern, *cedric.travelletti@stat.unibe.ch;
†david.ginsbourger@stat.unibe.ch

³Department of Mathematical Sciences, NTNU, jo.eidsvik@ntnu.no

⁴Underwater Systems and Technology Laboratory, Faculty of Engineering, University of Porto, kanna.rajan@fe.up.pt

Improving and optimizing oceanographic sampling is a crucial task for marine science and maritime resource management. Faced with limited resources in understanding processes in the water column, the combination of statistics and autonomous systems provides new opportunities for experimental design. In this work we develop efficient spatial sampling methods for characterizing regions, defined by simultaneous exceedances above prescribed thresholds of several responses, with an application focus on mapping coastal ocean phenomena based on temperature and salinity measurements. Specifically, we define a design criterion based on uncertainty in the excursions of vector-valued Gaussian random fields and derive tractable expressions for the expected integrated Bernoulli variance reduction in such a framework. We demonstrate how this criterion can be used to prioritize sampling efforts at locations that are ambiguous, making exploration more effective. We use simulations to study and compare properties of the considered approaches, followed by results from field deployments with an autonomous underwater vehicle as part of a study mapping the boundary of a river plume. The results demonstrate the potential of combining statistical methods and robotic platforms to effectively inform and execute data-driven environmental sampling.

REFERENCES

- AZZIMONTI, D., BECT, J., CHEVALIER, C. and GINSBOURGER, D. (2016). Quantifying uncertainties on excursion sets under a Gaussian random field prior. *SIAM/ASA J. Uncertain. Quantificat.* **4** 850–874. [MR3531737](https://doi.org/10.1137/141000749)
- AZZIMONTI, D., GINSBOURGER, D., CHEVALIER, C., BECT, J. and RICHT, Y. (2021). Adaptive design of experiments for Conservative estimation of excursion sets. *Technometrics* **63** 13–26. [MR4205688](https://doi.org/10.1080/00401706.2019.1693427)
- BECT, J., BACHOC, F. and GINSBOURGER, D. (2019). A supermartingale approach to Gaussian process based sequential design of experiments. *Bernoulli* **25** 2883–2919. [MR4003568](https://doi.org/10.3150/18-BEJ1074)
- BECT, J., GINSBOURGER, D., LI, L., PICHENY, V. and VAZQUEZ, E. (2012). Sequential design of computer experiments for the estimation of a probability of failure. *Stat. Comput.* **22** 773–793. [MR2909621](https://doi.org/10.1007/s11222-011-9241-4)
- BELLINGHAM, J. G. and RAJAN, K. (2007). Robotics in remote and hostile environments. *Science* **318** 1098–1102.
- BHATTACHARIYA, D., EIDSVIK, J. and MUKERJI, T. (2013). The value of information in portfolio problems with dependent projects. *Decis. Anal.* **10** 341–351. [MR3150459](https://doi.org/10.1287/deca.2013.0277)
- BINNEY, J., KRAUSE, A. and SUKHATME, G. S. (2013). Optimizing waypoints for monitoring spatiotemporal phenomena. *The International Journal of Robotics Research* **32** 873–888.
- BOLIN, D. and LINDGREN, F. (2015). Excursion and contour uncertainty regions for latent Gaussian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 85–106. [MR3299400](https://doi.org/10.1111/rssb.12055)

Key words and phrases. Excursion sets, Gaussian processes, experimental design, autonomous robots, ocean sampling, adaptive information gathering.

- BUESO, M., ANGULO, J. and ALONSO, F. (1998). A state-space model approach to optimum spatial sampling design based on entropy. *Environ. Ecol. Stat.* **5** 29–44.
- CHEVALIER, C., EMERY, X. and GINSBOURGER, D. (2015). Fast update of conditional simulation ensembles. *Math. Geosci.* **47** 771–789.
- CHEVALIER, C., GINSBOURGER, D. and EMERY, X. (2013). Corrected kriging update formulae for batch-sequential data assimilation. In *Mathematics of Planet Earth* (E. Pardo-Igúzquiza, C. Guardiola-Albert, J. Heredia, L. Moreno-Merino, J. Durán and J. Vargas-Guzmán, eds.). *Lecture Notes in Earth System Sciences*. Springer, Berlin.
- CHEVALIER, C., GINSBOURGER, D., BECT, J. and MOLCHANOV, I. (2013). Estimating and quantifying uncertainties on level sets using the Vorob'ev expectation and deviation with Gaussian process models. In *MODa 10 Advances in Model-Oriented Design and Analysis* (D. Uciński, A. C. Atkinson and C. Patan, eds.) Physica-Verlag, Heidelberg.
- CHEVALIER, C., GINSBOURGER, D., BECT, J., VAZQUEZ, E., PICHENY, V. and RICHEL, Y. (2014). Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics* **56** 455–465. [MR3290615 https://doi.org/10.1080/00401706.2013.860918](https://doi.org/10.1080/00401706.2013.860918)
- COSTA, M. J., PINTO, J., DIAS, P. S., PEREIRA, J., LIMA, K., RIBEIRO, M., SOUSA, J. B., LUKACZYK, T., MENDES, R. et al. (2018). Field report: Exploring fronts with multiple robots. In *2018 IEEE/OES Autonomous Underwater Vehicle Symposium*.
- DAS, J., PY, F., MAUGHAN, T., MESSIE, M., O'REILLY, T., RYAN, J., SUKHATME, G. S. and RAJAN, K. (2012). Coordinated sampling of dynamic oceanographic features with AUVs and drifters. *International Journal of Robotics Research* **31** 626–646.
- DAS, J., PY, F., HARVEY, J. B. J., RYAN, J. P., GELLEN, A., GRAHAM, R., CARON, D. A., RAJAN, K. and SUKHATME, G. S. (2015). Data-driven robotic sampling for marine ecosystem monitoring. *The International Journal of Robotics Research* **34** 1435–1452.
- DAVIDSON, F., ALVERA-AZCÁRATE, A., BARTH, A., BRASSINGTON, G. B., CHASSIGNET, E. P., CLEMENTI, E., DE MEY-FRÉMAUX, P., DIVAKARAN, P., HARRIS, C. et al. (2019). Synergies in operational oceanography: The intrinsic need for sustained ocean observations. *Front. Mar. Sci.* **6** 450.
- EIDSVIK, J., MUKERJI, T. and BHATTACHARJYA, D. (2015). *Value of Information in the Earth Sciences: Integrating Spatial Modeling and Decision Analysis*. Cambridge Univ. Press, Cambridge.
- FERREIRA, A. S., COSTA, M., PY, F., PINTO, J., SILVA, M. A., NIMMO-SMITH, A., JOHANSEN, T. A., DE SOUSA, J. B. and RAJAN, K. (2019). Advancing multi-vehicle deployments in oceanographic field experiments. *Autonomous Robots* **43** 1555–1574.
- FOSSUM, T. O., EIDSVIK, J., ELLINGSEN, I., ALVER, M. O., FRAGOSO, G. M., JOHNSEN, G., MENDES, R., LUDVIGSEN, M. and RAJAN, K. (2018). Information-driven robotic sampling in the coastal ocean. *J. Field Robot.* **35** 1101–1121.
- FOSSUM, T. O., FRAGOSO, G. M., DAVIES, E. J., ULLGREN, J. E., MENDES, R., JOHNSEN, G., ELLINGSEN, I., EIDSVIK, J., LUDVIGSEN, M. et al. (2019). Toward adaptive robotic sampling of phytoplankton in the coastal ocean. *Sci Robot* **4**. <https://doi.org/10.1126/scirobotics.aav3041>
- FOSSUM, T. O., TRAVELLETTI, C., EIDSVIK, J., GINSBOURGER, D. and RAJAN, K. (2021). Supplement to “Learning excursion sets of vector-valued Gaussian random fields for autonomous ocean sampling.” <https://doi.org/10.1214/21-AOAS1451SUPPA>
- FOSSUM, T. O., TRAVELLETTI, C., EIDSVIK, J., GINSBOURGER, D. and RAJAN, K. (2021). Source code for “Learning excursion sets of vector-valued Gaussian random fields for autonomous ocean sampling”—The MESLAS package. <https://doi.org/10.1214/21-AOAS1451SUPPB>
- FRENCH, J. P. and SAIN, S. R. (2013). Spatio-temporal exceedance locations and confidence regions. *Ann. Appl. Stat.* **7** 1421–1449. [MR3127953 https://doi.org/10.1214/13-AOAS631](https://doi.org/10.1214/13-AOAS631)
- GENTON, M. G. and KLEIBER, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statist. Sci.* **30** 147–163. [MR3353096 https://doi.org/10.1214/14-STS487](https://doi.org/10.1214/14-STS487)
- GENZ, A. and BRETZ, F. (2009). *Computation of Multivariate Normal and t Probabilities*. *Lecture Notes in Statistics* **195**. Springer, Dordrecht. [MR2840595 https://doi.org/10.1007/978-3-642-01689-9](https://doi.org/10.1007/978-3-642-01689-9)
- GINSBOURGER, D. (2018). *Wiley StatsRef: Statistics Reference Online. Sequential Design of Computer Experiments* 1–9.
- GNEITING, T., KLEIBER, W. and SCHLATHER, M. (2010). Matérn cross-covariance functions for multivariate random fields. *J. Amer. Statist. Assoc.* **105** 1167–1177. [MR2752612 https://doi.org/10.1198/jasa.2010.tm09420](https://doi.org/10.1198/jasa.2010.tm09420)
- GOTTLIEB, J., GRAHAM, R., MAUGHAN, T., PY, F., ELKAIM, G. and RAJAN, K. (2012). An experimental momentum-based front detection for autonomous underwater vehicles. In *IEEE International Conference on Robotics and Automation (ICRA)*.
- HOPKINS, J., LUCAS, M., DUFAU, C., SUTTON, M., STUM, J., LAURET, O. and CHANNELIERE, C. (2013). Detection and variability of the Congo river plume from satellite derived sea surface temperature, salinity, ocean colour and sea level. *Remote Sens. Environ.* **139** 365–385.

- KARAMAN, S. and FRAZZOLI, E. (2011). Sampling-based algorithms for optimal motion planning. *The International Journal of Robotics Research* **30** 846–894.
- KRIGE, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *J. of the Chem., Metal. and Mining Soc. of South Africa* **52** 119–139.
- LE GRATIET, L. and CANNAMELA, C. (2015). Cokriging-based sequential design strategies using fast cross-validation techniques for multi-fidelity computer codes. *Technometrics* **57** 418–427. [MR3384956](#) <https://doi.org/10.1080/00401706.2014.928233>
- LERMUSIAUX, P. F. J. (2006). Uncertainty estimation and prediction for interdisciplinary ocean dynamics. *J. Comput. Phys.* **217** 176–199. [MR2250531](#) <https://doi.org/10.1016/j.jcp.2006.02.010>
- LOW, K. H., DOLAN, J. M. and KHOSLA, P. K. (2009). Information-theoretic approach to efficient adaptive path planning for mobile robotic environmental sensing. *Proceedings of the International Conference on Automated Planning and Scheduling* 233–240.
- MELLUCCI, C., MENON, P. P., EDWARDS, C. and CHALLENGOR, P. (2018). Oceanic feature boundary mapping with an autonomous underwater glider. In *2018 Annual American Control Conference (ACC)* 5338–5343. IEEE, New York.
- MOLCHANOV, I. (2005). *Theory of Random Sets. Probability and Its Applications (New York)*. Springer London, Ltd., London. [MR2132405](#)
- MÜLLER, W. G. (2007). *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*, third revised and extended ed. *Contributions to Statistics*. Physica-Verlag, Heidelberg. [MR1795688](#)
- PINTO, J., MENDES, R., DA SILVA, J. C. B., DIAS, J. M. and DE SOUSA, J. B. (2018). Multiple autonomous vehicles applied to plume detection and tracking. In *2018 OCEANS—MTS/IEEE Kobe Techno-Oceans (OTO)* 1–6.
- POLOCZEK, M. U., WANG, J. and FRAZIER, P. I. (2017). Multi-information source optimization. In *Advances in Neural Information Processing Systems* 30.
- POWELL, W. B. (2016). Perspectives of approximate dynamic programming. *Ann. Oper. Res.* **241** 319–356. [MR3509421](#) <https://doi.org/10.1007/s10479-012-1077-6>
- PY, F., RAJAN, K. and MCGANN, C. (2010). A systematic agent framework for situated autonomous systems. In *9th International Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*.
- RAJAN, K. and PY, F. (2012). T-REX: Partitioned inference for AUV mission control. In *Further Advances in Unmanned Marine Vehicles* (G. N. Roberts and R. Sutton, eds.) The Institution of Engineering and Technology (IET).
- RAJAN, K., PY, F. and BERREIRO, J. (2012). Towards deliberative control in marine robotics. In *Marine Robot Autonomy* (M. Seto, ed.) Springer, Berlin.
- RICHARDSON, R. A. (2017). Sparsity in nonlinear dynamic spatiotemporal models using implied advection. *Environmetrics* **28** e2456. [MR3693639](#) <https://doi.org/10.1002/env.2456>
- RYAN, J. P., KUDELA, R. M., BIRCH, J. M., BLUM, M., BOWERS, H. A., CHAVEZ, F. P., DOUCETTE, G. J., HAYASHI, K., MARIN, R. III et al. (2017). Causality of an extreme harmful algal bloom in Monterey Bay, California, during the 2014–2016 northeast Pacific warm anomaly. *Geophys. Res. Lett.* **44** 5571–5579.
- SAHU, S. K. and CHALLENGOR, P. (2008). A space–time model for joint modeling of ocean temperature and salinity levels as measured by Argo floats. *Environmetrics* **19** 509–528. [MR2523911](#) <https://doi.org/10.1002/env.895>
- SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2003). *The Design and Analysis of Computer Experiments. Springer Series in Statistics*. Springer, New York. [MR2160708](#) <https://doi.org/10.1007/978-1-4757-3799-8>
- SIGRIST, F., KÜNSCH, H. R. and STAHEL, W. A. (2015). Stochastic partial differential equation based modelling of large space–time data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 3–33. [MR3299397](#) <https://doi.org/10.1111/rssb.12061>
- SINGH, A., KRAUSE, A., GUESTRIN, C. and KAISER, W. J. (2009). Efficient informative sensing using multiple robots. *J. Artificial Intelligence Res.* **34** 707–755. [MR2534376](#) <https://doi.org/10.1613/jair.2674>
- SMITH, R. N., PY, F., COOKSEY, P., SUKHATME, G. and RAJAN, K. (2014). Adaptive path planning for tracking ocean fronts with an autonomous underwater vehicle. In *Intl. Symp. on Experimental Robotics (ISER)*.
- SOSA, A., MADUREIRA, L., COELHO, J., PINTO, J., PEREIRA, J., SOSA, J. and DIAS, P. (2012). LAUV: The man-portable autonomous underwater vehicle. In *Navigation, Guidance and Control of Underwater Vehicles* 3 268–274.
- STEIN, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging. Springer Series in Statistics*. Springer, New York. [MR1697409](#) <https://doi.org/10.1007/978-1-4612-1494-6>
- STROH, R. (2018). Planification d’expériences numériques en multi-fidélité: Application à un simulateur d’incendies. Ph.D. thesis, Univ. Paris-Saclay.
- VARGAS-GUZMÁN, J. A. and JIM YEH, T. C. (1999). Sequential kriging and cokriging: Two powerful geostatistical approaches. *Stochastic Environmental Research and Risk Assessment* **13** 416–435.

- WACKERNAGEL, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Springer, Berlin.
- WIKLE, C. K., MILLIFF, R. F., HERBEL, R. and LEEDS, W. B. (2013). Modern statistical methods in oceanography: A hierarchical perspective. *Statist. Sci.* **28** 466–486. [MR3161583](#) <https://doi.org/10.1214/13-STS436>
- ZIDEK, J. V. and ZIMMERMAN, D. L. (2019). Monitoring network design. In *Handbook of Environmental and Ecological Statistics. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 499–522. CRC Press, Boca Raton, FL. [MR3889909](#)

AGGREGATED PAIRWISE CLASSIFICATION OF ELASTIC PLANAR SHAPES

BY MIN HO CHO^{*}, SEBASTIAN KURTEK[†] AND STEVEN N. MACEachern[‡]

Department of Statistics, Ohio State University, ^{*}cho.829@osu.edu; [†]kurtek.1@stat.osu.edu; [‡]snm@stat.osu.edu

The classification of shapes is of great interest in diverse areas ranging from medical imaging to computer vision and beyond. While many statistical frameworks have been developed for the classification problem, most are strongly tied to early formulations of the problem with an object to be classified described as a vector in a relatively low-dimensional Euclidean space. Statistical shape data have two main properties that suggest a need for a novel approach: (i) shapes are inherently infinite-dimensional with strong dependence among the positions of nearby points, and (ii) shape space is not Euclidean but is fundamentally curved. To accommodate these features of the data, we work with the square-root velocity function of the curves to provide a useful formal description of the shape, pass to tangent spaces of the manifold of shapes at projection points (which effectively separate shapes for pairwise classification in the training data) and use principal components within these tangent spaces to reduce dimensionality. We illustrate the impact of the projection point and choice of subspace on the misclassification rate with a novel method of combining pairwise classifiers.

REFERENCES

- BAI, X., LIU, W. and TU, Z. (2009). Integrating contour and skeleton for shape classification. In *IEEE International Conference on Computer Vision Workshops* 360–367.
- BERGER, J. O. (2013). *Statistical Decision Theory and Bayesian Analysis*, Springer Series in Statistics. Springer, New York. [MR0804611 https://doi.org/10.1007/978-1-4757-4286-2](https://doi.org/10.1007/978-1-4757-4286-2)
- BERGER, J. O. and PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122. [MR1394065 https://doi.org/10.2307/2291387](https://doi.org/10.2307/2291387)
- CHO, M. H., KURTEK, S. and MACEachern, S. N. (2021a). Supplement to “Aggregated pairwise classification of elastic planar shapes.” <https://doi.org/10.1214/21-AOAS1452SUPPA>
- CHO, M. H., KURTEK, S. and MACEachern, S. N. (2021b). Source code for “Aggregated pairwise classification of elastic planar shapes.” <https://doi.org/10.1214/21-AOAS1452SUPPB>
- COOTES, T. F., TAYLOR, C. J., COOPER, D. H. and GRAHAM, J. (1995). Active shape models—Their training and application. *Comput. Vis. Image Underst.* **61** 38–59.
- DRYDEN, I. L. and MARDIA, K. V. (1992). Size and shape analysis of landmark data. *Biometrika* **79** 57–68. [MR1158517 https://doi.org/10.1093/biomet/79.1.57](https://doi.org/10.1093/biomet/79.1.57)
- DRYDEN, I. L. and MARDIA, K. V. (2016). *Statistical Shape Analysis with Applications in R*, 2nd ed. Wiley Series in Probability and Statistics. Wiley, Chichester. [MR3559734 https://doi.org/10.1002/9781119072492](https://doi.org/10.1002/9781119072492)
- GLAUNÈS, J., QIU, A., MILLER, M. I. and YOUNES, L. (2008). Large deformation diffeomorphic metric curve mapping. *Int. J. Comput. Vis.* **80** 317–336.
- GRENANDER, U. and MILLER, M. I. (1998). Computational anatomy: An emerging discipline. *Quart. Appl. Math.* **56** 617–694. [MR1668732 https://doi.org/10.1090/qam/1668732](https://doi.org/10.1090/qam/1668732)
- GROVE, K. and KARCHER, H. (1973). How to conjugate C^1 -close group actions. *Math. Z.* **132** 11–20. [MR0356104 https://doi.org/10.1007/BF01214029](https://doi.org/10.1007/BF01214029)
- HOTZ, T., HUCKEMANN, S., MUNK, A., GAFFREY, D. and SLOBODA, B. (2010). Shape spaces for prealigned star-shaped objects—Studying the growth of plants by principal components analysis. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59** 127–143. [MR2750135 https://doi.org/10.1111/j.1467-9876.2009.00683.x](https://doi.org/10.1111/j.1467-9876.2009.00683.x)
- JOSHI, S. H., KLASSEN, E., SRIVASTAVA, A. and JERMYN, I. H. (2007). A novel representation for Riemannian analysis of elastic curves in $\mathbb{R}^n$. In *IEEE Conference on Computer Vision and Pattern Recognition* 1–7.

- KENDALL, D. G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *Bull. Lond. Math. Soc.* **16** 81–121. [MR0737237](#) <https://doi.org/10.1112/blms/16.2.81>
- KLASSEN, E., SRIVASTAVA, A., MIO, W. and JOSHI, S. H. (2004). Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 372–383.
- KURTEK, S., SRIVASTAVA, A., KLASSEN, E. and DING, Z. (2012). Statistical modeling of curves using shapes and related features. *J. Amer. Statist. Assoc.* **107** 1152–1165. [MR3010902](#) <https://doi.org/10.1080/01621459.2012.699770>
- KURTEK, S., SU, J., GRIMM, C., VAUGHAN, M., SOWELL, R. and SRIVASTAVA, A. (2013). Statistical analysis of manual segmentations of structures in medical images. *Comput. Vis. Image Underst.* **117** 1036–1050.
- LAGA, H., KURTEK, S., SRIVASTAVA, A., GOLZARIAN, M. and MIKLAVCIC, S. J. (2012). A Riemannian elastic metric for shape-based plant leaf classification. In *International Conference on Digital Image Computing Techniques and Applications* 1–7.
- LE, H. and KUME, A. (2000). Detection of shape changes in biological features. *J. Microsc.* **200** 140–147.
- MALLADI, R., SETHIAN, J. A. and VEMURI, B. C. (1996). A fast level set based algorithm for topology-independent shape modeling. *J. Math. Imaging Vision* **6** 269–289. [MR1390215](#) <https://doi.org/10.1007/BF00119843>
- MIO, W., SRIVASTAVA, A. and JOSHI, S. H. (2007). On shape of plane elastic curves. *Int. J. Comput. Vis.* **73** 307–324.
- PAL, S., WOODS, R. P., PANJIYAR, S., SOWELL, E. R., NARR, K. L. and JOSHI, S. H. (2017). A Riemannian framework for linear and quadratic discriminant analysis on the tangent space of shapes. In *Workshop on Differential Geometry in Computer Vision and Machine Learning* 726–734.
- PIZER, S. M., JUNG, S., GOSWAMI, D., VICORY, J., ZHAO, X., CHAUDHURI, R., DAMON, J. N., HUCKEMANN, S. and MARRON, J. S. (2013). Nested sphere statistics of skeletal models. In *Innovations for Shape Analysis: Models and Algorithms. Math. Vis.* 93–115. Springer, Heidelberg. [MR3075829](#) https://doi.org/10.1007/978-3-642-34141-0_5
- ROBINSON, D. T. (2012). Functional data analysis and partial shape matching in the square root velocity framework. Ph.D. thesis, The Florida State Univ. [MR3152393](#)
- SIDDIQI, K. and PIZER, S. M., eds. (2008). *Medial Representations: Mathematics, Algorithms and Applications. Computational Imaging and Vision* **37**. Springer, New York. [MR2547467](#) <https://doi.org/10.1007/978-1-4020-8658-8>
- SRIVASTAVA, A. and KLASSEN, E. P. (2016). *Functional and Shape Data Analysis. Springer Series in Statistics*. Springer, New York. [MR3821566](#)
- SRIVASTAVA, A., JAIN, A., JOSHI, S. and KAZISKA, D. (2006). Statistical shape models using elastic-string representations. In *Asian Conference on Computer Vision* 612–621. Springer, Berlin.
- SRIVASTAVA, A., KLASSEN, E., JOSHI, S. H. and JERMYN, I. H. (2011). Shape analysis of elastic curves in Euclidean spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 1415–1428.
- STRAIT, J., KURTEK, S., BARTHA, E. and MACEACHERN, S. N. (2017). Landmark-constrained elastic shape analysis of planar curves. *J. Amer. Statist. Assoc.* **112** 521–533. [MR3671749](#) <https://doi.org/10.1080/01621459.2016.1236726>
- VAILLANT, M., MILLER, M. I., YOUNES, L. and TROUVÉ, A. (2004). Statistics on diffeomorphisms via tangent space representations. *NeuroImage* **23** S161–S169.
- WEINBERGER, K. Q., BLITZER, J. and SAUL, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems* 1473–1480.
- WU, S. G., BAO, F. S., XU, E. Y., WANG, Y.-X., CHANG, Y.-F. and XIANG, Q.-L. (2007). A leaf recognition algorithm for plant classification using probabilistic neural network. In *IEEE International Symposium on Signal Processing and Information Technology* 11–16.
- YOUNES, L. (1998). Computable elastic distances between shapes. *SIAM J. Appl. Math.* **58** 565–586. [MR1617630](#) <https://doi.org/10.1137/S0036139995287685>
- YU, Q., MACEACHERN, S. N. and PERUGGIA, M. (2011). Bayesian synthesis: Combining subjective analyses, with an application to ozone data. *Ann. Appl. Stat.* **5** 1678–1698. [MR2849791](#) <https://doi.org/10.1214/10-AOAS444>
- ZAHN, C. T. and ROSKIES, R. Z. (1972). Fourier descriptors for plane closed curves. *IEEE Trans. Comput.* **21** 269–281. [MR0321383](#) <https://doi.org/10.1109/tc.1972.5008949>

A STATISTICAL PIPELINE FOR IDENTIFYING PHYSICAL FEATURES THAT DIFFERENTIATE CLASSES OF 3D SHAPES

BY BRUCE WANG¹, TIMOTHY SUDIJONO², HENRY KIRVESLAHTI³, TINGRAN GAO⁴, DOUGLAS M. BOYER⁵, SAYAN MUKHERJEE⁶ AND LORIN CRAWFORD⁷

¹Lewis-Sigler Institute for Integrative Genomics, Princeton University, bruce.wang@princeton.edu

²Division of Applied Mathematics, Brown University, timothy_sudijono@brown.edu

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

⁴Committee on Computational and Applied Mathematics, Department of Statistics, University of Chicago, gaotingran@gmail.com

⁵Department of Evolutionary Anthropology, Duke University, doug.boyer@duke.edu

⁶Department of Statistical Science, Department of Computer Science, Department of Mathematics, and Department of Bioinformatics & Biostatistics, Duke University, sayan@stat.duke.edu

⁷Microsoft Research New England, lcrawford@microsoft.com

The recent curation of large-scale databases with 3D surface scans of shapes has motivated the development of tools that better detect global patterns in morphological variation. Studies, which focus on identifying differences between shapes, have been limited to simple pairwise comparisons and rely on prespecified landmarks (that are often known). We present SINATRA, the first statistical pipeline for analyzing collections of shapes without requiring any correspondences. Our novel algorithm takes in two classes of shapes and highlights the physical features that best describe the variation between them. We use a rigorous simulation framework to assess our approach. Lastly, as a case study we use SINATRA to analyze mandibular molars from four different suborders of primates and demonstrate its ability recover known morphometric variation across phylogenies.

REFERENCES

- ADLER, D., NENADIC, O. and ZUCCHINI, W. (2003). RGL: An R-library for 3D visualization with OpenGL. In *Proceedings of the 35th Symposium of the Interface: Computing Science and Statistics, Salt Lake City* **35** 1–11.
- ANDERSON, J. T., WILLIS, J. H. and MITCHELL-OLDS, T. (2011). Evolutionary genetics of plant adaptation. *Trends Genet.* **27** 258–266.
- BELONGIE, S. (1999). Rodrigues' rotation formula. From MathWorld—A Wolfram Web Resource, created by Eric W. Weisstein. Available at <http://mathworld.wolfram.com/RodriguesRotationFormula.html>.
- BENDICH, P., MARRON, J. S., MILLER, E., PIELOCH, A. and SKWERER, S. (2016). Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.* **10** 198–218. MR3480493 <https://doi.org/10.1214/15-AOAS886>
- BOYER, D. M., LIPMAN, Y., CLAIR, E. S., PUENTE, J., PATEL, B. A., FUNKHOUSER, T., JERNVALL, J. and DAUBECHIES, I. (2011). Algorithms to automatically quantify the geometric similarity of anatomical surfaces. *Proc. Natl. Acad. Sci. USA* **108** 18221–18226. <https://doi.org/10.1073/pnas.1112822108>
- BOYER, D. M., PUENTE, J., GLADMAN, J. T., GLYNN, C., MUKHERJEE, S., YAPUNCICH, G. S. and DAUBECHIES, I. (2015). A new fully automated approach for aligning and comparing shapes. *Anat. Rec. (Hoboken)* **298** 249–276.
- BOYER, D. M., GUNNELL, G. F., KAUFMAN, S. and MCGEARY, T. M. (2016). Morphosource: Archiving and sharing 3-D digital specimen data. *The Paleontological Society Papers* **22** 157–181.
- CATES, J., ELHABIAN, S. and WHITAKER, R. (2017). Shapeworks: Particle-based shape correspondence and visualization software. In *Statistical Shape and Deformation Analysis* 257–298. Elsevier, Amsterdam.
- CHAUDHURI, A., KAKDE, D., SADEK, C., GONZALEZ, L. and KONG, S. (2017). The mean and median criteria for kernel bandwidth selection for support vector data description. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, 2017 842–849.

Key words and phrases. 3D image analysis, centrality measures, topological data analysis, Gaussian processes, evolutionary biology, phylogenetics.

- CHEN, J., KÄLLMAN, T., MA, X., GYLLENSTRAND, N., ZAINA, G., MORGANTE, M., BOUSQUET, J., ECKERT, A., WĘGRZYN, J. et al. (2012). Disentangling the roles of history and local selection in shaping clinal variation of allele frequencies and gene expression in Norway spruce (*Picea abies*). *Genetics* **191** 865–881.
- CHENG, L., RAMCHANDRAN, S., VATANEN, T., LIETZÉN, N., LAHESMAA, R., VEHTARI, A. and LÄHDESMÄKI, H. (2019). An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data. *Nat. Commun.* **10** 1798.
- COVER, T. and HART, P. (2006). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13** 21–27.
- CRAWFORD, L., ZENG, P., MUKHERJEE, S. and ZHOU, X. (2017). Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLoS Genet.* **13** e1006869.
- CRAWFORD, L., WOOD, K. C., ZHOU, X. and MUKHERJEE, S. (2018). Bayesian approximate Kernel regression with variable selection. *J. Amer. Statist. Assoc.* **113** 1710–1721. [MR3902240 https://doi.org/10.1080/01621459.2017.1361830](https://doi.org/10.1080/01621459.2017.1361830)
- CRAWFORD, L., FLAXMAN, S. R., RUNCIE, D. E. and WEST, M. (2019). Variable prioritization in non-linear black box methods: A genetic association case study. *Ann. Appl. Stat.* **13** 958–989. [MR3963559 https://doi.org/10.1214/18-AOAS1222](https://doi.org/10.1214/18-AOAS1222)
- CRAWFORD, L., MONOD, A., CHEN, A. X., MUKHERJEE, S. and RABADÁN, R. (2020). Predicting clinical outcomes in glioblastoma: An application of topological and functional data analysis. *J. Amer. Statist. Assoc.* **115** 1139–1150. [MR4143455 https://doi.org/10.1080/01621459.2019.1671198](https://doi.org/10.1080/01621459.2019.1671198)
- CROMPTON, R. H., SAVAGE, R. and SPEARS, I. R. (1998). The mechanics of food reduction in *Tarsius bancanus*. Hard-object feeder, soft-object feeder or both? *Folia Primatol. (Basel)* **69 Suppl 1** 41–59. <https://doi.org/10.1159/000052698>
- CURRY, J., MUKHERJEE, S. and TURNER, K. (2019). How many directions determine a shape and other sufficiency results for two topological transforms. Available at [arXiv:1805.09782](https://arxiv.org/abs/1805.09782).
- DUPUIS, P., GRENDER, U. and MILLER, M. I. (1998). Variational problems on flows of diffeomorphisms for image matching. *Quart. Appl. Math.* **56** 587–600. [MR1632326 https://doi.org/10.1090/qam/1632326](https://doi.org/10.1090/qam/1632326)
- FASY, B. T., MICKA, S., MILLMAN, D. L., SCHENFISCH, A. and WILLIAMS, L. (2018). Challenges in reconstructing shapes from Euler characteristic curves. Available at [arXiv:1811.11337](https://arxiv.org/abs/1811.11337).
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.
- GAO, T. (2015). Hypoelliptic diffusion maps and their applications in automated geometric morphometrics, Ph.D. thesis, Duke Univ.
- GAO, T. (2021). The diffusion geometry of fibre bundles: Horizontal diffusion maps. *Appl. Comput. Harmon. Anal.* **50** 147–215. [MR4162316 https://doi.org/10.1016/j.acha.2019.08.001](https://doi.org/10.1016/j.acha.2019.08.001)
- GAO, T., KOVALSKY, S. Z. and DAUBECHIES, I. (2019). Gaussian process landmarking on manifolds. *SIAM J. Math. Data Sci.* **1** 208–236. [MR3949707 https://doi.org/10.1137/18M1184035](https://doi.org/10.1137/18M1184035)
- GAO, T., KOVALSKY, S. Z., BOYER, D. M. and DAUBECHIES, I. (2019). Gaussian process landmarking for three-dimensional geometric morphometrics. *SIAM J. Math. Data Sci.* **1** 237–267. [MR3949708 https://doi.org/10.1137/18M1203481](https://doi.org/10.1137/18M1203481)
- GHRIST, R., LEVANGER, R. and MAI, H. (2018). Persistent homology and Euler integral transforms. *J. Appl. Comput. Topol.* **2** 55–60. [MR3873179 https://doi.org/10.1007/s41468-018-0017-1](https://doi.org/10.1007/s41468-018-0017-1)
- GIENAPP, P., TEPLITSKY, C., ALHO, J. S., MILLS, J. A. and MERILA, J. (2008). Climate change and evolution: Disentangling environmental and genetic responses. *Mol. Ecol.* **17** 167–178.
- GOPALAN, G. and BORNN, L. (2015). FastGP: An R package for Gaussian processes. Available at [arXiv:1507.06055](https://arxiv.org/abs/1507.06055).
- GOSWAMI, A. (2015). Phenome10K: A free online repository for 3-D scans of biological and palaeontological specimens. Website.
- GRAVEN, A. (1989). The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **326** 87–99.
- GUATELLI-STEINBERG, D. (2003). Primate dentition: An introduction to the teeth of non-human primates. *Am. J. Phys. Anthropol.* **121** 189–189.
- HECKERMAN, D., GURDASANI, D., KADIE, C., POMILLA, C., CARSTENSEN, T., MARTIN, H., EKORU, K., NSUBUGA, R. N., SSENOMO, G. et al. (2016). Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proc. Natl. Acad. Sci. USA* **113** 7377–7382.
- HENDERSON, C. R. (1984). *Applications of Linear Models in Animal Breeding*. Guelph, Ont.: University of Guelph. Includes index.
- HONG, Y., GOLLAND, P. and ZHANG, M. (2017). Fast geodesic regression for population-based image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 317–325. Springer, Berlin.
- HUANG, R., ACHLIOPTAS, P., GUIBAS, L. and OVSIANIKOV, M. (2019). Limit shapes—A tool for understanding shape differences and variability in 3D model collections. *Comput. Graph. Forum* **38** 187–202.

- JIANG, Y. and REIF, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* **201** 759–768.
- KANG, H. M., ZAITLEN, N. A., WADE, C. M., KIRBY, A., HECKERMAN, D., DALY, M. J. and ESKIN, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178** 1709–1723.
- KANG, H. M., SUL, J. H., SERVICE, S. K., ZAITLEN, N. A., KONG, S.-Y., FREIMER, N. B., SABATTI, C. and ESKIN, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42** 348–354.
- KENDALL, D. G. (1989). A survey of the statistical theory of shape. *Statist. Sci.* **4** 87–120. [MR1007558](#)
- LAI, Y.-T., YEUNG, C. K. L., OMLAND, K. E., PANG, E.-L., HAO, Y., LIAO, B.-Y., CAO, H.-F., ZHANG, B.-W., YEH, C.-F. et al. (2019). Standing genetic variation as the predominant source for adaptation of a songbird. *Proc. Natl. Acad. Sci. USA* **116** 2152–2157.
- LI, F., ZHANG, T., WANG, Q., GONZALEZ, M. Z., MARESH, E. L. and COAN, J. A. (2015). Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *Ann. Appl. Stat.* **9** 687–713. [MR3371331](#) <https://doi.org/10.1214/15-AOAS818>
- MCDONALD, K. R., BRODERICK, W. F., HUETTEL, S. A. and PEARSON, J. M. (2019). Bayesian nonparametric models characterize instantaneous strategies in a competitive dynamic game. *Nat. Commun.* **10** 1808.
- MILLER, E. (2015). Fruit flies and moduli: Interactions between biology and mathematics. *Notices Amer. Math. Soc.* **62** 1178–1184. [MR3408062](#) <https://doi.org/10.1090/noti1290>
- NEAL, R. M. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression and Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report No. 9702, Dept. of Statistics, Univ. Toronto.
- NEAL, R. M. (1999). Regression and classification using Gaussian process priors. In *Bayesian Statistics, 6 (Alcoceber, 1998)* 475–501. Oxford Univ. Press, New York. [MR1723510](#)
- NICKISCH, H. and RASMUSSEN, C. E. (2008). Approximations for binary Gaussian process classification. *J. Mach. Learn. Res.* **9** 2035–2078. [MR2452620](#)
- OUDOT, S. and SOLOMON, E. (2018). Inverse problems in topological persistence. Available at [arXiv:1810.10813](#).
- OVSJANIKOV, M., BEN-CHEN, M., SOLOMON, J., BUTSCHER, A. and GUIBAS, L. (2012). Functional maps: A flexible representation of maps between shapes. *ACM Trans. Graph.* **31** 30:1–30:11.
- PILLAI, N. S., WU, Q., LIANG, F., MUKHERJEE, S. and WOLPERT, R. L. (2007). Characterizing the function space for Bayesian kernel models. *J. Mach. Learn. Res.* **8** 1769–1797. [MR2332448](#)
- POZZI, L., HODGSON, J. A., BURELLA, A. S., RAAUMB, R. L. and DISOTELL, T. R. (2014). Primate phylogenetic relationships and divergence dates inferred from complete mitochondrial genomes. *Mol. Phylogenet. Evol.* **75** 165–183.
- PUENTE, J. (2013). Distances and algorithms to compare sets of shapes for automated biological morphometrics, Ph.D. thesis, Princeton Univ., Princeton, NJ.
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR2514435](#)
- RODRIGUEZ-NIEVA, J. F. and SCHEURER, M. S. (2019). Identifying topological order through unsupervised machine learning. *Nat. Phys.*
- 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** 1–12.
- SCHLAGER, S., ZHENG, G., LI, S. and SZÉKELY, G. (2017). Morpho and Rvcg—Shape analysis in R: R-packages for geometric morphometrics, shape analysis and surface manipulations. In *Statistical Shape and Deformation Analysis: Methods, Implementation and Applications* 217–256. Academic Press, San Diego.
- SCHLAGER, S., PROFICO, A., DI VINCENZO, F. and MANZI, G. (2018). Retrodeformation of fossil specimens based on 3D bilateral semi-landmarks: Implementation in the R package “Morpho”. *PLoS ONE* **13** e0194073.
- SCHÖLKOPF, B., HERBRICH, R. and SMOLA, A. J. (2001). A generalized representer theorem. In *Computational Learning Theory (Amsterdam, 2001). Lecture Notes in Computer Science* **2111** 416–426. Springer, Berlin. [MR2042050](#) https://doi.org/10.1007/3-540-44581-1_27
- SELA, M., AFLALO, Y. and KIMMEL, R. (2015). Computational caricaturization of surfaces. *Comput. Vis. Image Underst.* **141** 1–17.
- SELLKE, T., BAYARRI, M. J. and BERGER, J. O. (2001). Calibration of p values for testing precise null hypotheses. *Amer. Statist.* **55** 62–71. [MR1818723](#) <https://doi.org/10.1198/000313001300339950>
- SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. [MR3173712](#) <https://doi.org/10.1080/10618600.2012.681250>
- SINGLETON, K. R., CRAWFORD, L., TSUI, E., MANCHESTER, H. E., MAERTENS, O., LIU, X., LIBERTI, M. V., MAGPUSAO, A. N., STEIN, E. M. et al. (2017). Melanoma therapeutic strategies that select against resistance by exploiting MYC-driven evolutionary convergence. *Cell Rep.* **21** 2796–2812.

- ST CLAIR, E. M. and BOYER, D. M. (2016). Lower molar shape and size in prosimian and platyrrhine primates. *Am. J. Phys. Anthropol.* **161** 237–258.
- SWAIN, P. S., STEVENSON, K., LEARY, A., MONTANO-GUTIERREZ, L. F., CLARK, I. B. N., VOGEL, J. and PILIZOTA, T. (2016). Inferring time derivatives including cell growth rates using Gaussian processes. *Nat. Commun.* **7** 13766.
- TURNER, K., MUKHERJEE, S. and BOYER, D. M. (2014). Persistent homology transform for modeling shapes and surfaces. *Inf. Inference* **3** 310–344. [MR3311455 https://doi.org/10.1093/imaiai/iau011](https://doi.org/10.1093/imaiai/iau011)
- WANG, B., SUDIJONO, T., KIRVESLAHTI, H., GAO, T., BOYER, D. M., MUKHERJEE, S. and CRAWFORD, L. (2021). Supplement to “A statistical pipeline for identifying physical features that differentiate classes of 3D shapes.” <https://doi.org/10.1214/20-AOAS1430SUPPA>.
- WANG, B., SUDIJONO, T., KIRVESLAHTI, H., GAO, T., BOYER, D. M., MUKHERJEE, S. and CRAWFORD, L. (2021). Source Code for “A statistical pipeline for identifying physical features that differentiate classes of 3D shapes.” <https://doi.org/10.1214/20-AOAS1430SUPPB>.
- WILLIAMS, C. K. I. and BARBER, D. (1998). Bayesian classification with Gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **20** 1342–1351.
- WORSLEY, K. J. (1995). Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images. *Ann. Statist.* **23** 640–669. [MR1332586 https://doi.org/10.1214/aos/1176324540](https://doi.org/10.1214/aos/1176324540)
- YANG, Y. and ZOU, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Stat. Comput.* **25** 1129–1141. [MR3401877 https://doi.org/10.1007/s11222-014-9498-5](https://doi.org/10.1007/s11222-014-9498-5)
- ZHANG, Z., DAI, G. and JORDAN, M. I. (2011). Bayesian generalized kernel mixed models. *J. Mach. Learn. Res.* **12** 111–139. [MR2773550](https://doi.org/10.1214/11-AOS1093)
- ZHOU, X. and STEPHENS, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44** 821–825.
- ZHU, X. and STEPHENS, M. (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat. Commun.* **9** 4361.
- ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. [MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x)

SIMULTANEOUS INFERENCE OF PERIODS AND PERIOD-LUMINOSITY RELATIONS FOR MIRA VARIABLE STARS

BY SHIYUAN HE¹, ZHENFENG LIN², WENLONG YUAN³, LUCAS M. MACRI⁴ AND
JIANHUA Z. HUANG⁵

¹*Institute of Statistics and Big Data, Renmin University of China, heshiyuan@ruc.edu.cn*

²*Microsoft, zhenfeng.lin@microsoft.com*

³*Department of Physics and Astronomy, Johns Hopkins University, wyuan10@jhu.edu*

⁴*Department of Physics and Astronomy, Texas A&M University, lmacri@tamu.edu*

⁵*Department of Statistics, Texas A&M University, jianhua@stat.tamu.edu*

The period-luminosity relation (PLR) of Mira variable stars is an important tool to determine astronomical distances. The common approach of estimating the PLR is a two-step procedure that first estimates the Mira periods and then runs a linear regression of magnitude on log period. When the light curves are sparse and noisy, the accuracy of period estimation decreases and can suffer from aliasing effects. Some methods improve accuracy by incorporating complex model structures at the expense of significant computational costs. Another drawback of existing methods is that they only provide point estimation without proper estimation of uncertainty. To overcome these challenges, we develop a hierarchical Bayesian model that simultaneously models the quasi-periodic variations for a collection of Mira light curves while estimating their common PLR. By borrowing strengths through the PLR, our method automatically reduces the aliasing effect, improves the accuracy of period estimation and is capable of characterizing the estimation uncertainty. We develop a scalable stochastic variational inference algorithm for computation that can effectively deal with the multimodal posterior of period. The effectiveness of the proposed method is demonstrated through simulations and an application to observations of Miras in the Local Group galaxy M33. Without using ad hoc period correction tricks, our method achieves a distance estimate of M33 that is consistent with published work. Our method also shows superior robustness to downsampling of the light curves.

REFERENCES

- BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. [MR3671776 https://doi.org/10.1080/01621459.2017.1285773](https://doi.org/10.1080/01621459.2017.1285773)
- BONANOS, A. Z., STANEK, K. Z., KUDRITZKI, R. P., MACRI, L. M., SASSELOV, D. D., KALUZNY, J., STETSON, P. B., BERSIER, D., BRESOLIN, F. et al. (2006). The first direct distance determination to a detached eclipsing binary in m33. *Astrophys. J.* **652** 313.
- FLEWELLING, H. (2018). Pan-STARRS Data Release 2. In *American Astronomical Society Meeting Abstracts* 231 436.01.
- FREEDMAN, W. L., MADORE, B. F., GIBSON, B. K., FERRARESE, L., KELSON, D. D., SAKAI, S., MOULD, J. R., KENNICUTT, R. C. JR., FORD, H. C. et al. (2001). Final results from the hubble space telescope key project to measure the hubble constant. *Astrophys. J.* **553** 47.
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. [MR2221284 https://doi.org/10.1214/06-BA117A](https://doi.org/10.1214/06-BA117A)
- GHADIMI, S. and LAN, G. (2013). Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.* **23** 2341–2368. [MR3134439 https://doi.org/10.1137/120880811](https://doi.org/10.1137/120880811)
- GIEREN, W., GÓRSKI, M., PIETRZYŃSKI, G., KONORSKI, P., SUCHOMSKA, K., GRACZYK, D., PILECKI, B., BRESOLIN, F., KUDRITZKI, R.-P. et al. (2013). The araucaria project. A distance determination to the local group spiral m33 from near-infrared photometry of cepheid variables. *Astrophys. J.* **773** 69.

- HE, S., YUAN, W., HUANG, J. Z., LONG, J. and MACRI, L. M. (2016). Period estimation for sparsely sampled quasi-periodic light curves applied to miras. *Astron. J.* **152** 164.
- HE, S., LIN, Z., YUAN, W., MACRI, L. M. and HUANG, J. Z. (2021). Supplement to “Simultaneous inference of periods and period-luminosity relations for Mira variable stars.” <https://doi.org/10.1214/21-AOAS1440SUPPA>, <https://doi.org/10.1214/21-AOAS1440SUPPB>
- HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.* **14** 1303–1347. [MR3081926](https://arxiv.org/abs/1308.1926)
- HUANG, C. D., RIESS, A. G., HOFFMANN, S. L., KLEIN, C., BLOOM, J., YUAN, W., MACRI, L. M., JONES, D. O., WHITELOCK, P. A. et al. (2018). A near-infrared period–luminosity relation for miras in ngc 4258, an anchor for a new distance ladder. *Astrophys. J.* **857** 67.
- JAVADI, A., SABERI, M., VAN LOON, J. T., KHOSROSHAHI, H., GOLABATOONI, N. and MIRTORABI, M. T. (2015). The uk infrared telescope m33 monitoring project-IV. Variable red giant stars across the galactic disc. *Mon. Not. R. Astron. Soc.* **447** 3973–3991.
- JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.
- LOMB, N. R. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophys. Space Sci.* **39** 447–462.
- LONG, J. P., CHI, E. C. and BARANIUK, R. G. (2016). Estimating a common period for a set of irregularly sampled functions with applications to periodic variable star data. *Ann. Appl. Stat.* **10** 165–197. [MR3480492](https://arxiv.org/abs/1508.0885) <https://doi.org/10.1214/15-AOAS885>
- MACRI, L. M. M. (2001). The extragalactic distance scale. Ph.D. thesis, Harvard Univ.
- MACRI, L., STANEK, K., SASSELOV, D., KROCKENBERGER, M. and KALUZNY, J. (2001). Direct distances to nearby galaxies using detached eclipsing binaries and cepheids. VI. Variables in the central part of m33. *Astron. J.* **121** 870.
- MOWLAVI, N., LECOEUR-TAÏBI, I., LEBZELTER, T., RIMOLDINI, L., LORENZ, D., AUDARD, M., DE RIDDER, J., EYER, L., GUY, L. et al. (2018). Gaia data release 2-the first gaia catalogue of long-period variable candidates. *Astron. Astrophys.* **618** A58.
- PELLERIN, A. and MACRI, L. M. (2011). The m 33 synoptic stellar survey. I. Cepheid variables. *Astrophys. J., Suppl. Ser.* **193** 26.
- PIETRZYŃSKI, G., GIEREN, W., GRACZYK, D., THOMPSON, I., PILECKI, B., NARDETTO, N., KUDRITZKI, R.-P., BRESOLIN, F., BONO, G. et al. (2013). A precise and accurate distance to the large magellanic cloud from late-type eclipsing-binary systems. In *Advancing the Physics of Cosmic Distances* (R. de Grijs, ed.). *IAU Symposium* **289** 169–172.
- RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR2514435](https://arxiv.org/abs/2514435)
- RIESS, A. G., MACRI, L., CASERTANO, S., LAMPEITL, H., FERGUSON, H. C., FILIPPENKO, A. V., JHA, S. W., LI, W. and CHORNOCK, R. (2011). A 3% solution: Determination of the hubble constant with the hubble space telescope and wide field camera 3. *Astrophys. J.* **730** 119.
- RIESS, A. G., MACRI, L. M., HOFFMANN, S. L., SCOLNIC, D., CASERTANO, S., FILIPPENKO, A. V., TUCKER, B. E., REID, M. J., JONES, D. O. et al. (2016). A 2.4% determination of the local value of the hubble constant. *Astrophys. J.* **826** 56.
- RIESS, A. G., CASERTANO, S., YUAN, W., MACRI, L. M. and SCOLNIC, D. (2019). Large magellanic cloud cepheid standards provide a 1% foundation for the determination of the hubble constant and stronger evidence for physics beyond Λ cdm. *Astrophys. J.* **876** 85.
- SARAJEDINI, A., BARKER, M., GEISLER, D., HARDING, P. and SCHOMMER, R. (2006). Rr lyrae variables in m33. I. Evidence for a field halo population. *Astron. J.* **132** 1361.
- SCARGLE, J. D. (1982). December. Studies in astronomical time series analysis. II—Statistical aspects of spectral analysis of unevenly spaced data. *Astrophys. J.* **263** 835–853.
- SCOWCROFT, V., BERSIER, D., MOULD, J. and WOOD, P. R. (2009). The effect of metallicity on cepheid magnitudes and the distance to m33. *Mon. Not. R. Astron. Soc.* **396** 1287–1296.
- SOSZYŃSKI, I., UDALSKI, A., SZYMAŃSKI, M. K., KUBIAK, M., PIETRZYŃSKI, G., WYRZYKOWSKI, Ł., SZEWCZYK, O., ULACZYK, K. and POLESKI, R. (2009). The optical gravitational lensing experiment. The OGLE-III catalog of variable stars. III. RR lyrae stars in the large magellanic cloud. *Acta Astronomica* **59** 1–18.
- UDALSKI, A., SZYMAŃSKI, M. K., SOSZYŃSKI, I. and POLESKI, R. (2008). The optical gravitational lensing experiment. Final reductions of the ogle-iii data. *Acta Astronomica* **58** 69–87.
- VANDERPLAS, J. T. and IVEZIC, Ž. (2015). Periodograms for multiband astronomical time series. *Astrophys. J.* **812** 18.
- WAINWRIGHT, M. J., JORDAN, M. I. et al. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

- WANG, Y. and BLEI, D. M. (2019). Frequentist consistency of variational Bayes. *J. Amer. Statist. Assoc.* **114** 1147–1161. [MR4011769 https://doi.org/10.1080/01621459.2018.1473776](https://doi.org/10.1080/01621459.2018.1473776)
- WHITELOCK, P. A. and FEAST, M. W. (2014). Gaia, variable stars and the distance scale. *EAS Publications Series* **67** 263–269.
- WHITELOCK, P., MENZIES, J., FEAST, M., NSENGIYUMVA, F. and MATSUNAGA, N. (2014). VizieR online data catalog: Jhks photometry of agb stars in ngc 6822 (whitelock+, 2013). *VizieR Online Data Catalog* **742**.
- YUAN, W., MACRI, L. M., HE, S., HUANG, J. Z., KANBUR, S. M. and NGEOW, C.-C. (2017). Large magellanic cloud near-infrared synoptic survey. V. Period–luminosity relations of miras. *Astron. J.* **154** 149.
- YUAN, W., MACRI, L. M., JAVADI, A., LIN, Z. and HUANG, J. Z. (2018). Near-infrared mira period–luminosity relations in m33. *Astron. J.* **156** 112.
- ZECHMEISTER, M. and KÜRSTER, M. (2009). The generalised lomb-scargle periodogram. A new formalism for the floating-mean and Keplerian periodograms. *Astron. Astrophys.* **496** 577–584.
- ZHANG, C., BUTEPAGE, J., KJELLSTROM, H. and MANDT, S. (2019). Advances in variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.* **41** 2008–2026.

SCALABLE PENALIZED SPATIOTEMPORAL LAND-USE REGRESSION FOR GROUND-LEVEL NITROGEN DIOXIDE

BY KYLE P. MESSIER¹ AND MATTHIAS KATZFUSS²

¹*Division of the National Toxicology Program, National Institute of Environmental Health Sciences, kyle.messier@nih.gov*

²*Department of Statistics, Texas A&M University, katzfuss@gmail.com*

Nitrogen dioxide (NO₂) is a primary constituent of traffic-related air pollution and has well-established harmful environmental and human-health impacts. Knowledge of the spatiotemporal distribution of NO₂ is critical for exposure and risk assessment. A common approach for assessing air pollution exposure is linear regression involving spatially referenced covariates, known as land-use regression (LUR). We develop a scalable approach for simultaneous variable selection and estimation of LUR models with spatiotemporally correlated errors, by combining a general-Vecchia Gaussian-process approximation with a penalty on the LUR coefficients. In comparison to existing methods using simulated data, our approach resulted in higher model-selection specificity and sensitivity and in better prediction in terms of calibration and sharpness, for a wide range of relevant settings. In our spatiotemporal analysis of daily, US-wide, ground-level NO₂ data, our approach was more accurate, and produced a sparser and more interpretable model. Our daily predictions elucidate spatiotemporal patterns of NO₂ concentrations across the United States, including significant variations between cities and intra-urban variation. Thus, our predictions will be useful for epidemiological and risk-assessment studies seeking daily, national-scale predictions, and they can be used in acute-outcome health-risk assessments.

REFERENCES

- ABATZOGLOU, J. T., RUPP, D. E. and MOTE, P. W. (2014). Seasonal climate variability and change in the Pacific Northwest of the United States. *J. Climate* **27** 2125–2142.
- ALEXEEFF, S. E., ROY, A., SHAN, J., LIU, X., MESSIER, K., APTE, J. S., PORTIER, C., SIDNEY, S. and VAN DEN EEDEN, S. K. (2018). High-resolution mapping of traffic related air pollution with Google Street View cars and incidence of cardiovascular events within neighborhoods in Oakland, CA. *Environ. Health* **17** 1–13. <https://doi.org/10.1186/s12940-018-0382-1>
- APTE, J. S., MESSIER, K. P., GANI, S., BRAUER, M., KIRCHSTETTER, T. W., LUNDEN, M. M., MARSHALL, J. D., PORTIER, C. J., VERMEULEN, R. C. et al. (2017). High-resolution air pollution mapping with Google street view cars: Exploiting big data. *Environ. Sci. Technol.* **51** 6999–7008.
- APTE, J., GANI, S., CHAMBLISS, S., MESSIER, K., LUNDEN, M. et al. (2019). Potential underestimation of ultrafine particle exposure when using proxy pollutants: Lessons from long-term measurements at fixed sites and mobile monitoring. *Environ. Epidemiol.* **3** 13–14.
- BANERJEE, M., CAPOZZOLI, M., MCSWEENEY, L. and SINHA, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canad. J. Statist.* **27** 3–23. [MR1703616 https://doi.org/10.2307/3315487](https://doi.org/10.2307/3315487)
- BECKERMAN, B. S., JERRETT, M., SERRE, M. L., MARTIN, R. V., LEE, S., DONKELAAR, A. V., ROSS, Z., SU, J. and BURNETT, R. T. (2013). A hybrid approach to estimating national scale spatiotemporal variability of PM_{2.5} in the contiguous United States. *Environ. Sci. Technol.* **47** 7233–7241. <https://doi.org/10.1021/es400039u>
- BOERSMA, K. F., ESKES, H. J., VEEFKIND, J. P., BRINKSMA, E. J., VAN DER A, R. J., SNEEP, M., VAN DEN OORD, G. H. J., LEVELT, P. F., STAMMES, P. et al. (2007). Near-real time retrieval of tropospheric NO₂ from OMI. *Atmos. Chem. Phys.* **7** 2103–2118.
- BREHENY, P. and HUANG, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5** 232–253. [MR2810396 https://doi.org/10.1214/10-AOAS388](https://doi.org/10.1214/10-AOAS388)

Key words and phrases. General Vecchia approximation, spatial statistics, Gaussian process, variable selection, air pollution, Kriging.

- BRIGGS, D. J., COLLINS, S., ELLIOTT, P., FISCHER, P., KINGHAM, S., LEBRET, E., PRYL, K., VAN REEUWIJK, H., SMALLBONE, K. et al. (1997). Mapping urban air pollution using GIS: A regression-based approach. *Int. J. Geogr. Inf. Sci.* **11** 699–718. <https://doi.org/10.1080/136588197242158>
- CENTER FOR INTERNATIONAL EARTH SCIENCE INFORMATION NETWORK—CIESIN—COLUMBIA UNIVERSITY (2018). Gridded population of the world, version 4 (GPWv4): Population density, revision 11 [data set]. <https://doi.org/10.7927/H49C6VHW>
- CHATTERJEE, A. and LAHIRI, S. N. (2011). Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.* **106** 608–625. [MR2847974 https://doi.org/10.1198/jasa.2011.tm10159](https://doi.org/10.1198/jasa.2011.tm10159)
- COULLETTE, A. D., MONEY, E. S., SERRE, M. L. and NOBLE, R. T. (2009). Space/time analysis of fecal pollution and rainfall in an eastern North Carolina estuary. *Environ. Sci. Technol.* **43** 3728–35.
- DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016a). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. [MR3538706 https://doi.org/10.1080/01621459.2015.1044091](https://doi.org/10.1080/01621459.2015.1044091)
- DATTA, A., BANERJEE, S., FINLEY, A. O., HAMM, N. A. S. and SCHAAP, M. (2016b). Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *Ann. Appl. Stat.* **10** 1286–1316. [MR3553225 https://doi.org/10.1214/16-AOAS931](https://doi.org/10.1214/16-AOAS931)
- DE HOOGH, K., CHEN, J., GULLIVER, J., HOFFMANN, B., HERTEL, O., KETZEL, M., BAUWELINCK, M., VAN DONKELAAR, A., HVIDTFELDT, U. A. et al. (2018). Spatial PM_{2.5}, NO₂, O₃, and BC models for Western Europe—evaluation of spatiotemporal stability. *Environ. Int.* **120** 81–92. <https://doi.org/10.1016/j.envint.2018.07.036>
- DE HOOGH, K., SAUCY, A., SHTEIN, A., SCHWARTZ, J., WEST, E. A., STRASSMANN, A., PUHAN, M., ROOSLI, M., STAFOGGIA, M. et al. (2019). Predicting fine-scale daily NO₂ for 2005–2016 incorporating OMI satellite data across Switzerland. *Environ. Sci. Technol.* **53** 10279–10287.
- EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. With discussion, and a rejoinder by the authors. [MR2060166 https://doi.org/10.1214/009053604000000067](https://doi.org/10.1214/009053604000000067)
- FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. [MR1946581 https://doi.org/10.1198/016214501753382273](https://doi.org/10.1198/016214501753382273)
- 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 https://doi.org/10.1016/j.csda.2008.09.008](https://doi.org/10.1016/j.csda.2008.09.008)
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.
- GAUDERMAN, W. J., AVOL, E., LURMANN, F., KUENZLI, N., GILLILAND, F., PETERS, J. and MC-CONNELL, R. (2005). Childhood asthma and exposure to traffic and nitrogen dioxide. *Epidemiology* 737–743.
- GNEITING, T. and KATZFUSS, M. (2014). Probabilistic forecasting. *Annu. Rev. Stat. Appl.* **1** 125–151. <https://doi.org/10.1146/annurev-statistics-062713-085831>
- GRUNDSTRÖM, M., HAK, C., CHEN, D., HALLQUIST, M. and PLEIJEL, H. (2015). Variation and co-variation of PM₁₀, particle number concentration, NO_x and NO₂ in the urban air-relationships with wind speed, vertical temperature gradient and weather type. *Atmos. Environ.* **120** 317–327.
- GUAN, Y., JOHNSON, M. C., KATZFUSS, M., MANNSHARDT, E., MESSIER, K. P., REICH, B. J. and SONG, J. J. (2020). Fine-scale spatiotemporal air pollution analysis using mobile monitors on Google Street View vehicles. *J. Amer. Statist. Assoc.* **115** 1111–1124. [MR4143453 https://doi.org/10.1080/01621459.2019.1665526](https://doi.org/10.1080/01621459.2019.1665526)
- GUINNESS, J. (2018). Permutation and grouping methods for sharpening Gaussian process approximations. *Technometrics* **60** 415–429. [MR3878098 https://doi.org/10.1080/00401706.2018.1437476](https://doi.org/10.1080/00401706.2018.1437476)
- HASTIE, T., TIBSHIRANI, R. and TIBSHIRANI, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint. Available at [arXiv:1707.08692](https://arxiv.org/abs/1707.08692).
- HEATON, M. J., DATTA, A., FINLEY, A. O. et al. (2019). A case study competition among methods for analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* **24** 398–425. [MR3996451 https://doi.org/10.1007/s13253-018-00348-w](https://doi.org/10.1007/s13253-018-00348-w)
- HENDERSON, S. B., BECKERMAN, B., JERRETT, M. and BRAUER, M. (2007). Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environ. Sci. Technol.* **41** 2422–2428. <https://doi.org/10.1021/es0606780>
- HOEK, G., BELEN, R., DE HOOGH, K., VIENNEAU, D., GULLIVER, J., FISCHER, P. and BRIGGS, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* **42** 7561–7578. <https://doi.org/10.1016/j.atmosenv.2008.05.057>
- HOEK, G., KRISHNAN, R. M., BELEN, R., PETERS, A., OSTRO, B., BRUNEKREEF, B. and KAUFMAN, J. D. (2013). Long-term air pollution exposure and cardio-respiratory mortality: A review. *Environ. Health* **12** 43.

- HOLCOMB, D. A., MESSIER, K. P., SERRE, M. L., ROWNY, J. G. and STEWART, J. R. (2018). Geostatistical prediction of microbial water quality throughout a stream network using meteorology, land cover, and spatiotemporal autocorrelation. *Environ. Sci. Technol.* **52**. <https://doi.org/10.1021/acs.est.8b01178>
- HOMER, C., DEWITZ, J., YANG, L., JIN, S., DANIELSON, P., XIAN, G., COULSTON, J., HEROLD, N., WICKHAM, J. et al. (2015). Completion of the 2011 National Land Cover Database for the conterminous United States—representing a decade of land cover change information. *Photogramm. Eng. Remote Sens.* **81** 345–354.
- HUGHES, J. and HARAN, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 139–159. MR3008275 <https://doi.org/10.1111/j.1467-9868.2012.01041.x>
- JUREK, M. and KATZFUSS, M. (2018). Multi-resolution filters for massive spatio-temporal data. Available at [arXiv:1810.04200](https://arxiv.org/abs/1810.04200).
- KATZFUSS, M. (2017). A multi-resolution approximation for massive spatial datasets. *J. Amer. Statist. Assoc.* **112** 201–214. MR3646566 <https://doi.org/10.1080/01621459.2015.1123632>
- KATZFUSS, M. and GONG, W. (2020). A class of multi-resolution approximations for large spatial datasets. *Statist. Sinica* **30** 2203–2226. <https://doi.org/10.1007/s13253-020-00401-7>
- KATZFUSS, M. and GUINNESS, J. (2021). A general framework for Vecchia approximations of Gaussian processes. *Statist. Sci.* **36** 124–141. MR4194207 <https://doi.org/10.1214/19-STS755>
- KATZFUSS, M., GUINNESS, J., GONG, W. and ZILBER, D. (2020a). Vecchia approximations of Gaussian-process predictions. *J. Agric. Biol. Environ. Stat.* **25** 383–414. MR4139037 <https://doi.org/10.1007/s13253-020-00401-7>
- KATZFUSS, M., JUREK, M., ZILBER, D., GONG, W., GUINNESS, J., ZHANG, J. and SCHAEFER, F. (2020b). GPvecchia: Fast Gaussian-process inference using Vecchia approximations. R package version 0.1.3.
- KERCKHOFFS, J., HOEK, G., PORTENGEN, L., BRUNEKREEF, B. and VERMEULEN, R. C. (2019). Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. *Environ. Sci. Technol.* **53** 1413–1421.
- KNIBBS, L. D., HEWSON, M. G., BECHLE, M. J., MARSHALL, J. D. and BARNETT, A. G. (2014). A national satellite-based land-use regression model for air pollution exposure assessment in Australia. *Environ. Res.* **135** 204–211. <https://doi.org/10.1016/j.envres.2014.09.011>
- LARKIN, A., GEDDES, J. A., MARTIN, R. V., XIAO, Q., LIU, Y., MARSHALL, J. D., BRAUER, M. and HYSTAD, P. (2017). Global land use regression model for nitrogen dioxide air pollution. *Environ. Sci. Technol.* **51** 6957–6964.
- LI, R. and SUDJANTO, A. (2005). Analysis of computer experiments using penalized likelihood in Gaussian kriging models. *Technometrics* **47** 111–120. MR2188073 <https://doi.org/10.1198/004017004000000671>
- LI, J. J., FALTINGS, B., SAUKH, O., HASENFRATZ, D. and BEUTEL, J. (2012). Sensing the air we breathe—the OpenSense Zurich dataset. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.
- LIU, H., ONG, Y.-S., SHEN, X. and CAI, J. (2020). When Gaussian process meets big data: A review of scalable GPs. *IEEE Trans. Neural Netw. Learn. Syst.* **31** 4405–4423. MR4169962
- MAUZERALL, D. L., SULTAN, B., KIM, N. and BRADFORD, D. F. (2005). NO_x emissions from large point sources: Variability in ozone production, resulting health damages and economic costs. *Atmos. Environ.* **39** 2851–2866.
- MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). *p*-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681. MR2750584 <https://doi.org/10.1198/jasa.2009.tm08647>
- MESSIER, K. P., AKITA, Y. and SERRE, M. L. (2012). Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. *Environ. Sci. Technol.* **46** 2772–80. <https://doi.org/10.1021/es203152a>
- MESSIER, K. P. and KATZFUSS, M. (2021a). Supplement to scalable penalized spatiotemporal land-use regression for ground-level nitrogen dioxide. *Ann. Appl. Stat.* <https://doi.org/10.1214/20-AOAS1422SUPPA>
- MESSIER, K. P. and KATZFUSS, M. (2021b). R-code for scalable penalized spatiotemporal land-use regression for ground-level nitrogen dioxide. *Ann. Appl. Stat.* <https://doi.org/10.1214/20-AOAS1422SUPPB>
- MESSIER, K. P., KANE, E., BOLICH, R. and SERRE, M. L. (2014). Nitrate variability in groundwater of North Carolina using monitoring and private well data models. *Environ. Sci. Technol.* **48**. <https://doi.org/10.1021/es502725f>
- MESSIER, K. P., CAMPBELL, T., BRADLEY, P. J. and SERRE, M. L. (2015). Estimation of groundwater Radon in North Carolina using land use regression and Bayesian maximum entropy. *Environ. Sci. Technol.* **49** 9817–9825. <https://doi.org/10.1021/acs.est.5b01503>
- MILLER, D. J., ACTKINSON, B., PADILLA, L., GRIFFIN, R. J., MOORE, K., LEWIS, P. G. T., GARDNER-FROLICK, R., CRAFT, E., PORTIER, C. J. et al. (2020). Characterizing elevated urban air pollutant spatial patterns with mobile monitoring in Houston, Texas. *Environ. Sci. Technol.*

- MILLS, I. C., ATKINSON, R. W., KANG, S., WALTON, H. and ANDERSON, H. R. (2015). Quantitative systematic review of the associations between short-term exposure to nitrogen dioxide and mortality and hospital admissions. *BMJ Open* **5** e006946. <https://doi.org/10.1136/bmjopen-2014-006946>
- MOORE, D. K., JERRETT, M., MACK, W. J. and KÜNZLI, N. (2007). A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA. *J. Environ. Monit.* **9** 246–252. <https://doi.org/10.1039/b615795e>
- NASA/METI/AIST/JAPAN SPACESYSTEMS and U.S./JAPAN ASTER SCIENCE TEAM (2019). ASTER global digital elevation model V003 [data set]. <https://doi.org/10.5067/ASTER/ASTGTM.003>
- NOVOTNY, E. V., BECHLE, M. J., MILLET, D. B. and MARSHALL, J. D. (2011). National satellite-based land-use regression: NO₂ in the United States. *Environ. Sci. Technol.* **45** 4407–4414.
- PADRÓ-MARTÍNEZ, L. T., PATTON, A. P., TRULL, J. B., ZAMORE, W., BRUGGE, D. and DURANT, J. L. (2012). Mobile monitoring of particle number concentration and other traffic-related air pollutants in a near-highway neighborhood over the course of a year. *Atmos. Environ.* **61** 253–264.
- REYES, J. M. and SERRE, M. L. (2014). An LUR/BME framework to estimate PM_{2.5} explained by on road, mobile and stationary sources. *Environ. Sci. Technol.* **48** 1736–44. <https://doi.org/10.1021/es4040528>
- ROSENLUND, M., BERGLIND, N., PERSHAGEN, G., HALLQVIST, J., JONSON, T. and BELLANDER, T. (2006). Long-term exposure to urban air pollution and myocardial infarction. *Epidemiology* 383–390.
- ROSENLUND, M., PICCIOTTO, S., FORASTIERE, F., STAFOGGIA, M. and PERUCCI, C. A. (2008). Traffic-related air pollution in relation to incidence and prognosis of coronary heart disease. *Epidemiology* 121–128.
- ROSENLUND, M., BELLANDER, T., NORDQUIST, T. and ALFREDSSON, L. (2009). Traffic-generated air pollution and myocardial infarction. *Epidemiology* 265–271.
- ROSS, Z., ITO, K., JOHNSON, S., YEE, M., PEZESHKI, G., CLOUGHERTY, J. E., SAVITZ, D. and MATTE, T. (2013). Spatial and temporal estimation of air pollutants in New York City: Exposure assignment for use in a birth outcomes study. *Environ. Health* **12** 51.
- SAMPSON, P. D., RICHARDS, M., SZPIRO, A. A., BERGEN, S., SHEPPARD, L., LARSON, T. V. and KAUFMAN, J. D. (2013). A regionalized national universal kriging model using partial least squares regression for estimating annual PM_{2.5} concentrations in epidemiology. *Atmos. Environ.* **75** 383–392. <https://doi.org/10.1016/j.atmosenv.2013.04.015>
- SANG, H., JUN, M. and HUANG, J. Z. (2011). Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. *Ann. Appl. Stat.* **5** 2519–2548. <https://doi.org/10.1214/11-AOAS478>
- SCHÄFER, F., KATZFUSS, M. and OWHADI, H. (2020). Sparse Cholesky factorization by Kullback–Leibler minimization. Available at [arXiv:2004.14455](https://arxiv.org/abs/2004.14455).
- SCHINDLER, D. W. (1988). Effects of acid rain on freshwater ecosystems. *Science* **239** 149–157.
- SEINFELD, J. H. and PANDIS, S. N. (2016). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. Wiley, New York.
- SNELSON, E. and GHAHRAMANI, Z. (2007). Local and global sparse Gaussian process approximations. In *Artificial Intelligence and Statistics* 11 (AISTATS).
- SU, J. G., JERRETT, M. and BECKERMAN, B. (2009). A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. *Sci. Total Environ.* **407** 3890–3898. <https://doi.org/10.1016/j.scitotenv.2009.01.061>
- TADONO, T., ISHIDA, H., ODA, F., NAITO, S., MINAKAWA, K. and IWAMOTO, H. (2014). Precise global DEM generation by ALOS PRISM. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* **2** 71.
- TANG, W., ZHANG, L. and BANERJEE, S. (2019). On identifiability and consistency of the nugget in Gaussian spatial process models. Available at [arXiv:1908.05726](https://arxiv.org/abs/1908.05726).
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. <https://doi.org/10.2307/2346178>
- US ENVIRONMENTAL PROTECTION AGENCY (1999). U.S. EPA technical bulletin: Nitrogen oxides (NO_x), why and how they are controlled.
- US ENVIRONMENTAL PROTECTION AGENCY (2016). Integrated science assessment for oxides of nitrogen (final report). Technical report, EPA/600/R-15/068, US Environmental Protection Agency, National Center for Environmental Assessment Research, Research Triangle Park, NC.
- US ENVIRONMENTAL PROTECTION AGENCY (2017). 2017 national emissions inventory.
- US ENVIRONMENTAL PROTECTION AGENCY (2019). Air quality system pre-generated data files. Available at <https://www.epa.gov/outdoor-air-quality-data/download-daily-data>.
- VECCHIA, A. V. (1988). Estimation and model identification for continuous spatial processes. *J. Roy. Statist. Soc. Ser. B* **50** 297–312. <https://doi.org/10.2307/2346178>
- VOLK, H. E., LURMANN, F., PENFOLD, B., HERTZ-PICCIOTTO, I. and MCCONNELL, R. (2013). Traffic-related air pollution, particulate matter, and autism. *JAMA Psychiatr.* **70** 71–77. <https://doi.org/10.1001/jamapsychiatry.2013.266>

- WEISS, D. J., NELSON, A., GIBSON, H. S., TEMPERLEY, W., PEEDELL, S., LIEBER, A., HANCHER, M., POYART, E., BELCHIOR, S. et al. (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* **553** 333.
- WU, H., WANG, C. and WU, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14** 232–243. <https://doi.org/10.1093/biostatistics/kxs033>
- XIE, Y., XU, L., DENG, X., HONG, Y., KOLIVRAS, K. and GAINES, D. N. (2019). Spatial variable selection and an application to Virginia Lyme disease emergence. *J. Amer. Statist. Assoc.* **114** 1466–1480. [MR4047274 https://doi.org/10.1080/01621459.2018.1564670](https://doi.org/10.1080/01621459.2018.1564670)
- XU, X., HA, S. U. and BASNET, R. (2016). A review of epidemiological research on adverse neurological effects of exposure to ambient air pollution. *Front Public Health* **4** 157. <https://doi.org/10.3389/fpubh.2016.00157>
- YOUNG, M. T., BECHLE, M. J., SAMPSON, P. D., SZPIRO, A. A., MARSHALL, J. D., SHEPPARD, L. and KAUFMAN, J. D. (2016). Satellite-based NO₂ and model validation in a national prediction model based on universal kriging and land-use regression. *Environ. Sci. Technol.* **50** 3686–3694. <https://doi.org/10.1021/acs.est.5b05099>
- ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261. [MR2054303 https://doi.org/10.1198/016214504000000241](https://doi.org/10.1198/016214504000000241)
- ZILBER, D. and KATZFUSS, M. (2021). Vecchia–Laplace approximations of generalized Gaussian processes for big non-Gaussian spatial data. *Comput. Statist. Data Anal.* **153** 107081. [MR4146817 https://doi.org/10.1016/j.csda.2020.107081](https://doi.org/10.1016/j.csda.2020.107081)
- ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. [MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x)

PROBABILISTIC FORECASTING OF THE ARCTIC SEA ICE EDGE WITH CONTOUR MODELING

BY HANNAH M. DIRECTOR¹, ADRIAN E. RAFTERY² AND CECILIA M. BITZ³

¹*Department of Statistics, University of Washington, direch@uw.edu*

²*Departments of Statistics and Sociology, University of Washington, raftery@uw.edu*

³*Department of Atmospheric Sciences, University of Washington, bitz@uw.edu*

Sea ice, or frozen ocean water, freezes and melts every year in the Arctic. Forecasts of where sea ice will be located weeks to months in advance have become more important as the amount of sea ice declines due to climate change, for maritime planning and other uses. Typical sea ice forecasts are made with ensemble models, physics-based models of sea ice and the surrounding ocean and atmosphere. This paper introduces Mixture Contour Forecasting, a method to forecast sea ice probabilistically using a mixture of two distributions, one based on postprocessed output from ensembles and the other on observed sea ice patterns in recent years. At short lead times, these forecasts are better calibrated than unadjusted dynamic ensemble forecasts and other statistical reference forecasts. To produce these forecasts, a statistical technique is introduced that directly models the sea ice edge contour, the boundary around the region that is ice-covered. Mixture Contour Forecasting and reference methods are evaluated for monthly sea ice forecasts for 2008–2016 at lead times ranging from 0.5–6.5 months using one of the European Centre for Medium-Range Weather Forecasts ensembles.

REFERENCES

- BIVAND, R. and RUNDEL, C. (2020). rgeos: Interface to geometry engine—open source (‘GEOS’). R package version 0.5-5.
- BLANCHARD-WRIGGLESWORTH, E., CULLATHER, R. I., WANG, W., ZHANG, J. and BITZ, C. M. (2015). Model forecast skill and sensitivity to initial conditions in the seasonal sea ice outlook. *Geophys. Res. Lett.* **42** 8042–8048.
- BOLIN, D. and LINDGREN, F. (2015). Excursion and contour uncertainty regions for latent Gaussian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 85–106. [MR3299400 https://doi.org/10.1111/rssb.12055](https://doi.org/10.1111/rssb.12055)
- BRIER, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78** 1–3.
- BUSHUK, M., MSADEK, R., WINTON, M., VECCHI, G. A., GUDGEL, R., ROSATI, A. and YANG, X. (2017). Skillful regional prediction of Arctic sea ice on seasonal timescales. *Geophys. Res. Lett.* **44** 4953–4964.
- CAVALIERI, D. J. and PARKINSON, C. L. (2012). Arctic sea ice variability and trends, 1979–2010. *Cryosphere* **6** 881.
- CHEVALLIER, M., SALAS Y MÉLIA, D., VOLDOIRE, A., DÉQUÉ, M. and GARRIC, G. (2013). Seasonal forecasts of the pan-Arctic sea ice extent using a GCM-based seasonal prediction system. *J. Climate* **26** 6092–6104.
- COMISO, J. (2017). Bootstrap sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS. Version 3. NASA National Snow and Ice Data Center Distributed Active Archive Center, Boulder, Colorado, USA.
- COMISO, J. C., PARKINSON, C. L., GERSTEN, R. and STOCK, L. (2008). Accelerated decline in the Arctic sea ice cover. *Geophys. Res. Lett.* **35**.
- COPERNICUS CLIMATE CHANGE SERVICE (2019). Copernicus climate change service climate data store. Available at <https://cds.climate.copernicus.eu>.
- 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. With discussion. [MR0501537](https://doi.org/10.1111/rssb.12055)
- DIRECTOR, H. M., RAFTERY, A. E. and BITZ, C. M. (2017). Improved sea ice forecasting through spatiotemporal bias correction. *J. Climate* **30** 9493–9510.

- DIRECTOR, H. M., RAFTERY, A. E. and BITZ, C. M. (2021). Supplement to “Probabilistic forecasting of the Arctic sea ice edge with contour modeling.” <https://doi.org/10.1214/20-AOAS1405SUPPA>, <https://doi.org/10.1214/20-AOAS1405SUPPB>
- DIRKSON, A., DENIS, B. and MERRYFIELD, W. J. (2019). A multimodel approach for improving seasonal probabilistic forecasts of regional Arctic sea ice. *Geophys. Res. Lett.*
- DIRKSON, A., MERRYFIELD, W. J. and MONAHAN, A. H. (2019). Calibrated probabilistic forecasts of Arctic sea ice concentration. *J. Climate* **32** 1251–1271.
- DOUGLAS, D. H. and PEUCKER, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica* **10** 112–122.
- EDDELBUEITTEL, D., FRANÇOIS, R., ALLAIRE, J., USHEY, K., KOU, Q., RUSSEL, N., CHAMBERS, J. and BATES, D. (2011). rcpp: Seamless R and C++ integration. *J. Stat. Softw.* **40** 1–18.
- ELLIS, B. and BRIGHAM, L. (2009). Arctic marine shipping assessment 2009 report.
- EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS (2017). ECMWF SEAS5 user guide. Available at https://www.ecmwf.int/sites/default/files/media/library/2017-10/System5_guide.pdf.
- FRENCH, J. P. and HOETING, J. A. (2016). Credible regions for exceedance sets of geostatistical data. *Environmetrics* **27** 4–14. MR3457233 <https://doi.org/10.1002/env.2371>
- GNEITING, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19** 1327–1349. MR3102554 <https://doi.org/10.3150/12-BEJSP06>
- 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 <https://doi.org/10.1111/j.1467-9868.2007.00587.x>
- GUEMAS, V., BLANCHARD-WRIGGLESWORTH, E., CHEVALLIER, M., DAY, J. J., DÉQUÉ, M., DOBLAS-REYES, F. J., FUČKAR, N. S., GERME, A., HAWKINS, E. et al. (2016). A review on Arctic sea-ice predictability and prediction on seasonal to decadal time-scales. *Q. J. R. Meteorol. Soc.* **142** 546–561.
- HOLLAND, M. M., BAILEY, D. A. and VAVRUS, S. (2011). Inherent sea ice predictability in the rapidly changing Arctic environment of the community climate system model, version 3. *Clim. Dyn.* **36** 1239–1253.
- HUBER, P. J. (2011). *Robust Statistics*. Springer, Berlin.
- JOHNSON, S. J., STOCKDALE, T. N., FERRANTI, L., BALMASADA, M. A., MOLTENI, F., MAGNUSSON, L., TIETSCHKE, S., DECREMER, D., WEISHEIMER, A., BALSAMO, G. et al. (2019). SEAS5: The new ECMWF seasonal forecast system. *Geosci. Model Dev.* **12**.
- MELIA, N., HAINES, K. and HAWKINS, E. (2016). Sea ice decline and 21st century trans-Arctic shipping routes. *Geophys. Res. Lett.* **43** 9720–9728.
- MSADEK, R., VECCHI, G. A., WINTON, M. and GUDGEL, R. G. (2014). Importance of initial conditions in seasonal predictions of Arctic sea ice extent. *Geophys. Res. Lett.* **41** 5208–5215.
- NATIONAL SNOW & ICE DATA CENTER (2017). Region mask for the Northern Hemisphere. Available at http://nsidc.org/data/polar-stereo/tools_masks.html.
- PEBESMA, E. J. and BIVAND, R. S. (2005). Classes and methods for spatial data in R. *R News* **5** 9–13.
- RAFTERY, A. E., GNEITING, T., BALABDAOUI, F. and POLAKOWSKI, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133** 1155–1174.
- SIGMOND, M., FYFE, J. C., FLATO, G. M., KHARIN, V. V. and MERRYFIELD, W. J. (2013). Seasonal forecast skill of Arctic sea ice area in a dynamical forecast system. *Geophys. Res. Lett.* **40** 529–534.
- SMITH, L. C. and STEPHENSON, S. R. (2013). New trans-Arctic shipping routes navigable by midcentury. *Proc. Natl. Acad. Sci. USA* **110** E1191–E1195.
- STROEVE, J. C., SERREZE, M. C., HOLLAND, M. M., KAY, J. E., MALANIK, J. and BARRETT, A. P. (2012). The Arctic’s rapidly shrinking sea ice cover: A research synthesis. *Clim. Change* **110** 1005–1027.
- TIETSCHKE, S., DAY, J. J., GUEMAS, V., HURLIN, W. J., KEELEY, S. P. E., MATEI, D., MSADEK, R., COLLINS, M. and HAWKINS, E. (2014). Seasonal to interannual Arctic sea ice predictability in current global climate models. *Geophys. Res. Lett.* **41** 1035–1043.
- WANG, W., CHEN, M. and KUMAR, A. (2013). Seasonal prediction of Arctic sea ice extent from a coupled dynamical forecast system. *Mon. Weather Rev.* **141** 1375–1394.
- WAYAND, N. E., BITZ, C. M. and BLANCHARD-WRIGGLESWORTH, E. (2019). A year-round subseasonal-to-seasonal sea ice prediction portal. *Geophys. Res. Lett.* **46** 3298–3307.
- WICKHAM, H., AVERICK, M., BRYAN, J., CHANG, W., D’AGOSTINO MCGOWAN, L., FRANÇOIS, R., GROLEMUND, G., HAYES, A., HENRY, L. et al. (2019). Welcome to the tidyverse. *J. Open Sour. Softw.* **4** 1686.
- ZAMPIERI, L., GOESSLING, H. F. and JUNG, T. (2018). Bright prospects for Arctic sea ice prediction on sub-seasonal time scales. *Geophys. Res. Lett.* **45** 9731–9738.
- ZHANG, B. and CRESSIE, N. (2019). Estimating spatial changes over time of Arctic sea ice using hidden 2×2 tables. *J. Time Series Anal.* **40** 288–311. MR3946154 <https://doi.org/10.1111/jtsa.12425>

- ZHANG, B. and CRESSIE, N. (2020). Bayesian inference of spatio-temporal changes of Arctic sea ice. *Bayesian Anal.* **15** 605–631. [MR4097811](#) <https://doi.org/10.1214/20-BA1209>
- ZHUANG, J. (2018). xESMF: Universal regridding for geospatial data.
- ZIMMERMAN, D. L. and STEIN, M. (2010). Classical geostatistical methods. In *Handbook of Spatial Statistics. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 29–44. CRC Press, Boca Raton, FL. [MR2730951](#) <https://doi.org/10.1201/9781420072884-c3>

ADDITIVE STACKING FOR DISAGGREGATE ELECTRICITY DEMAND FORECASTING

BY CHRISTIAN CAPEZZA^{1,*}, BIAGIO PALUMBO^{1,†}, YANNIG GOUDE²,
SIMON N. WOOD³ AND MATTEO FASIOLO⁴

¹Department of Industrial Engineering, University of Naples Federico II, * christian.capezza@unina.it;
† biagio.palumbo@unina.it

²Électricité de France R&D, yannig.goude@edf.fr

³School of Mathematics, University of Edinburgh, simon.wood@ed.ac.uk

⁴School of Mathematics, University of Bristol, matteo.fasiolo@bristol.ac.uk

Future grid management systems will coordinate distributed production and storage resources to manage, in a cost effective fashion, the increased load and variability brought by the electrification of transportation and by a higher share of weather dependent production. Electricity demand forecasts at a low level of aggregation will be key inputs for such systems. We focus on forecasting demand at the individual household level, which is more challenging than forecasting aggregate demand, due to the lower signal-to-noise ratio and to the heterogeneity of consumption patterns across households. We propose a new ensemble method for probabilistic forecasting which borrows strength across the households while accommodating their individual idiosyncrasies. In particular, we develop a set of models or “experts” which capture different demand dynamics, and we fit each of them to the data from each household. Then, we construct an aggregation of experts where the ensemble weights are estimated on the whole data set, the main innovation being that we let the weights vary with the covariates by adopting an additive model structure. In particular, the proposed aggregation method is an extension of regression stacking where the mixture weights are modelled using linear combinations of parametric, smooth or random effects. The methods for building and fitting additive stacking models are implemented by the `gamFactory` R package, available at <https://github.com/mfasiolo/gamFactory>.

REFERENCES

- APLEY, D. W. and ZHU, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 1059–1086. [MR4136503](#)
- BASSETTI, F., CASARIN, R. and RAVAZZOLO, F. (2018). Bayesian nonparametric calibration and combination of predictive distributions. *J. Amer. Statist. Assoc.* **113** 675–685. [MR3832218](#) <https://doi.org/10.1080/01621459.2016.1273117>
- BREIMAN, L. (1996). Stacked regressions. *Mach. Learn.* **24** 49–64.
- CAPEZZA, C., PALUMBO, B., GOUDE, Y., WOOD, S. N. and FASIOLO, M. (2021a). Supplement to “Additive stacking for disaggregate electricity demand forecasting.” <https://doi.org/10.1214/20-AOAS1417SUPPA>.
- CAPEZZA, C., PALUMBO, B., GOUDE, Y., WOOD, S. N. and FASIOLO, M. (2021b). Supplement to “Additive stacking for disaggregate electricity demand forecasting.” <https://doi.org/10.1214/20-AOAS1417SUPPB>.
- CESA-BIANCHI, N. and LUGOSI, G. (2006). *Prediction, Learning, and Games*. Cambridge Univ. Press, Cambridge. [MR2409394](#) <https://doi.org/10.1017/CBO9780511546921>
- COMMISSION FOR ENERGY REGULATION (2012). CER Smart Metering Project—Electricity Customer Behaviour Trial, 2009–2010 [dataset]. 1st Edition. Irish Social Science Data Archive. SN: 0012-00. Available at www.ucd.ie/issda/CER-electricity.
- COSCRATO, V., DE ALMEIDA INÁCIO, M. H. and IZBICKI, R. (2020). The NN-stacking: Feature weighted linear stacking through neural networks. *Neurocomputing* **399** 141–152.

Key words and phrases. Electricity demand forecasting, probabilistic forecast, regression stacking, ensemble methods, generalised additive models.

- DEVAINE, M., GAILLARD, P., GOUDE, Y. and STOLTZ, G. (2013). Forecasting electricity consumption by aggregating specialized experts: A review of the sequential aggregation of specialized experts, with an application to Slovakian and French country-wide one-day-ahead (half-)hourly predictions. *Mach. Learn.* **90** 231–260. [MR3015743](#) <https://doi.org/10.1007/s10994-012-5314-7>
- FASIOLO, M., NEDELLEC, R., GOUDE, Y. and WOOD, S. N. (2020). Scalable visualization methods for modern generalized additive models. *J. Comput. Graph. Statist.* **29** 78–86. [MR4085865](#) <https://doi.org/10.1080/10618600.2019.1629942>
- FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. [MR1873328](#) <https://doi.org/10.1214/aos/1013203451>
- GNEITING, T. and RANJAN, R. (2013). Combining predictive distributions. *Electron. J. Stat.* **7** 1747–1782. [MR3080409](#) <https://doi.org/10.1214/13-EJS823>
- HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. CRC Press, London. [MR1082147](#)
- KOENKER, R. and BASSETT, G. JR. (1978). Regression quantiles. *Econometrica* **46** 33–50. [MR0474644](#) <https://doi.org/10.2307/1913643>
- LEBLANC, M. and TIBSHIRANI, R. (1996). Combining estimates in regression and classification. *J. Amer. Statist. Assoc.* **91** 1641–1650. [MR1439105](#) <https://doi.org/10.2307/2291591>
- MCALINN, K. and WEST, M. (2019). Dynamic Bayesian predictive synthesis in time series forecasting. *J. Econometrics* **210** 155–169. [MR3944768](#) <https://doi.org/10.1016/j.jeconom.2018.11.010>
- MORAL-BENITO, E. (2015). Model averaging in economics: An overview. *J. Econ. Surv.* **29** 46–75.
- NYCHKA, D. (1988). Bayesian confidence intervals for smoothing splines. *J. Amer. Statist. Assoc.* **83** 1134–1143. [MR0997592](#)
- PIMM, A. J., COCKERILL, T. T. and TAYLOR, P. G. (2018). Time-of-use and time-of-export tariffs for home batteries: Effects on low voltage distribution networks. *Journal of Energy Storage* **18** 447–458.
- PUDJANTO, D., DJAPIC, P., AUNEDI, M., GAN, C. K., STRBAC, G., HUANG, S. and INFELD, D. (2013). Smart control for minimizing distribution network reinforcement cost due to electrification. *Energy Policy* **52** 76–84.
- RANJAN, R. and GNEITING, T. (2010). Combining probability forecasts. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 71–91. [MR2751244](#) <https://doi.org/10.1111/j.1467-9868.2009.00726.x>
- RIGBY, R. A. and STASINOPOULOS, D. M. (2005). Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 507–554. [MR2137253](#) <https://doi.org/10.1111/j.1467-9876.2005.00510.x>
- SILL, J., TAKÁCS, G., MACKEY, L. and LIN, D. (2009). Feature-weighted linear stacking. Preprint. Available at [arXiv:0911.0460](#).
- SILVERMAN, B. W. (1986). *Density Estimation*. CRC Press, Boca Raton.
- STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36** 111–147. [MR0356377](#)
- WOLPERT, D. H. (1992). Stacked generalization. *Neural Netw.* **5** 241–259.
- WOOD, S. N. (2017). *Generalized Additive Models. Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. [MR3726911](#)
- WOOD, S. N., PYA, N. and SÄFKEN, B. (2016). Smoothing parameter and model selection for general smooth models. *J. Amer. Statist. Assoc.* **111** 1548–1563. [MR3601714](#) <https://doi.org/10.1080/01621459.2016.1180986>
- YAO, Y., VEHTARI, A., SIMPSON, D. and GELMAN, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Anal.* **13** 917–1003. [MR3853125](#) <https://doi.org/10.1214/17-BA1091>

CAUSAL MEDIATION ANALYSIS FOR SPARSE AND IRREGULAR LONGITUDINAL DATA

BY SHUXI ZENG^{1,*}, STACY ROSENBAUM^{2,4}, SUSAN C. ALBERTS³,
 ELIZABETH A. ARCHIE⁴ AND FAN LI^{1,†}

¹*Department of Statistical Science, Duke University, *zengshx777@gmail.com; †fl35@duke.edu*

²*Department of Anthropology, University of Michigan, rosenbas@umich.edu*

³*Departments of Biology and Evolutionary Anthropology, Duke University, alberts@duke.edu*

⁴*Department of Biological Sciences, University of Notre Dame, earchie@nd.edu*

Causal mediation analysis seeks to investigate how the treatment effect of an exposure on outcomes is mediated through intermediate variables. Although many applications involve longitudinal data, the existing methods are not directly applicable to settings where the mediator and outcome are measured on sparse and irregular time grids. We extend the existing causal mediation framework from a functional data analysis perspective, viewing the sparse and irregular longitudinal data as realizations of underlying smooth stochastic processes. We define causal estimands of direct and indirect effects accordingly and provide corresponding identification assumptions. For estimation and inference, we employ a functional principal component analysis approach for dimension reduction and use the first few functional principal components instead of the whole trajectories in the structural equation models. We adopt the Bayesian paradigm to accurately quantify the uncertainties. The operating characteristics of the proposed methods are examined via simulations. We apply the proposed methods to a longitudinal data set from a wild baboon population in Kenya to investigate the causal relationships between early adversity, strength of social bonds between animals and adult glucocorticoid hormone concentrations. We find that early adversity has a significant direct effect (a 9–14% increase) on females' glucocorticoid concentrations across adulthood but find little evidence that these effects were mediated by weak social bonds.

REFERENCES

- ALBERTS, S. C. and ALTMANN, J. (2012). The Amboseli baboon research project: 40 years of continuity and change. In *Long-Term Field Studies of Primates* 261–287. Springer, Berlin.
- ANDERSON, M. and MARMOT, M. (2011). The effects of promotions on heart disease: Evidence from Whitehall. *Econ. J.* **122** 555–589.
- BARON, R. M. and KENNY, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51** 1173–1182. <https://doi.org/10.1037//0022-3514.51.6.1173>
- BATESON, P., BARKER, D., CLUTTON-BROCK, T., DEB, D., D'UDINE, B., FOLEY, R. A., GLUCKMAN, P., GODFREY, K., KIRKWOOD, T. et al. (2004). Developmental plasticity and human health. *Nature* **430** 419.
- BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. [MR2806429 https://doi.org/10.1093/biomet/asr013](https://doi.org/10.1093/biomet/asr013)
- BIND, M.-A. C., VANDERWEELE, T. J., COULL, B. A. and SCHWARTZ, J. D. (2016). Causal mediation analysis for longitudinal data with exogenous exposure. *Biostatistics* **17** 122–134. [MR3449855 https://doi.org/10.1093/biostatistics/kxv029](https://doi.org/10.1093/biostatistics/kxv029)
- CASE, A. and PAXSON, C. (2011). The long reach of childhood health and circumstance: Evidence from the Whitehall II study. *Econ. J.* **121** F183–F204.
- COHEN, S. and WILLS, T. A. (1985). Stress, social support, and the buffering hypothesis. *Psychol. Bull.* **98** 310–357.

Key words and phrases. Causal inference, functional principal component analysis, mediation, longitudinal data, sparse and irregular data.

- DANIELS, M. J., ROY, J. A., KIM, C., HOGAN, J. W. and PERRI, M. G. (2012). Bayesian inference for the causal effect of mediation. *Biometrics* **68** 1028–1036. [MR3040009](#) <https://doi.org/10.1111/j.1541-0420.2012.01781.x>
- DIDELEZ, V. (2019). Defining causal mediation with a longitudinal mediator and a survival outcome. *Lifetime Data Anal.* **25** 593–610. [MR4015373](#) <https://doi.org/10.1007/s10985-018-9449-0>
- DIDELEZ, V., DAWID, A. P. and GENELETTI, S. (2006). Direct and indirect effects of sequential treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* 138–146.
- DURANTE, D. (2017). A note on the multiplicative gamma process. *Statist. Probab. Lett.* **122** 198–204. [MR3584158](#) <https://doi.org/10.1016/j.spl.2016.11.014>
- FORASTIERE, L., MATTEI, A. and DING, P. (2018). Principal ignorability in mediation analysis: Through and beyond sequential ignorability. *Biometrika* **105** 979–986. [MR3877878](#) <https://doi.org/10.1093/biomet/asy053>
- GLUCKMAN, P. D., HANSON, M. A., COOPER, C. and THORNBURG, K. L. (2008). Effect of in utero and early-life conditions on adult health and disease. *N. Engl. J. Med.* **359** 61–73.
- GOLDSMITH, J., GREVEN, S. and CRAINICEANU, C. (2013). Corrected confidence bands for functional data using principal components. *Biometrics* **69** 41–51. [MR3058050](#) <https://doi.org/10.1111/j.1541-0420.2012.01808.x>
- HAN, K., HADJIPANTELOS, P. Z., WANG, J.-L., KRAMER, M. S., YANG, S., MARTIN, R. M. and MÜLLER, H.-G. (2018). Functional principal component analysis for identifying multivariate patterns and archetypes of growth, and their association with long-term cognitive development. *PLoS ONE* **13** e0207073.
- HOLT-LUNSTAD, J., SMITH, T. B. and LAYTON, J. B. (2010). Social relationships and mortality risk: A meta-analytic review. *PLoS Med.* **7** e1000316. <https://doi.org/10.1371/journal.pmed.1000316>
- IMAI, K., KEELE, L. and TINGLEY, D. (2010). A general approach to causal mediation analysis. *Psychol. Methods* **15** 309–334. <https://doi.org/10.1037/a0020761>
- IMAI, K., KEELE, L. and YAMAMOTO, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* **25** 51–71. [MR2741814](#) <https://doi.org/10.1214/10-STS321>
- IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge Univ. Press, Cambridge.
- JIANG, C.-R. and WANG, J.-L. (2010). Covariate adjusted functional principal components analysis for longitudinal data. *Ann. Statist.* **38** 1194–1226. [MR2604710](#) <https://doi.org/10.1214/09-AOS742>
- JIANG, C.-R. and WANG, J.-L. (2011). Functional single index models for longitudinal data. *Ann. Statist.* **39** 362–388. [MR2797850](#) <https://doi.org/10.1214/10-AOS845>
- KIM, C., DANIELS, M. J., MARCUS, B. H. and ROY, J. A. (2017). A framework for Bayesian nonparametric inference for causal effects of mediation. *Biometrics* **73** 401–409. [MR3665957](#) <https://doi.org/10.1111/biom.12575>
- KIM, C., DANIELS, M., LI, Y., MILBURY, K. and COHEN, L. (2018). A Bayesian semiparametric latent variable approach to causal mediation. *Stat. Med.* **37** 1149–1161. [MR3777965](#) <https://doi.org/10.1002/sim.7572>
- KIM, C., DANIELS, M. J., HOGAN, J. W., CHOIRAT, C. and ZIGLER, C. M. (2019). Bayesian methods for multiple mediators: Relating principal stratification and causal mediation in the analysis of power plant emission controls. *Ann. Appl. Stat.* **13** 1927–1956. [MR4019162](#) <https://doi.org/10.1214/19-AOAS1260>
- KOWAL, D. R. and BOURGEOIS, D. C. (2020). Bayesian function-on-scalars regression for high-dimensional data. *J. Comput. Graph. Statist.* **29** 629–638. [MR4153187](#) <https://doi.org/10.1080/10618600.2019.1710837>
- LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 963–974.
- LANDERMAN, R., GEORGE, L. K., CAMPBELL, R. T. and BLAZER, D. G. (1989). Alternative models of the stress buffering hypothesis. *Am. J. Community Psychol.* **17** 625–642.
- LANGE, T., VANSTEELANDT, S. and BEKAERT, M. (2012). A simple unified approach for estimating natural direct and indirect effects. *Am. J. Epidemiol.* **176** 190–195.
- LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. [MR0836430](#) <https://doi.org/10.1093/biomet/73.1.13>
- LIN, S.-H., YOUNG, J., LOGAN, R., TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2017a). Parametric mediational *g*-formula approach to mediation analysis with time-varying exposures, mediators, and confounders. *Epidemiology* **28** 266.
- LIN, S.-H., YOUNG, J. G., LOGAN, R. and VANDERWEELE, T. J. (2017b). Mediation analysis for a survival outcome with time-varying exposures, mediators, and confounders. *Stat. Med.* **36** 4153–4166. [MR3713656](#) <https://doi.org/10.1002/sim.7426>
- LINDQUIST, M. A. (2012). Functional causal mediation analysis with an application to brain connectivity. *J. Amer. Statist. Assoc.* **107** 1297–1309. [MR3036396](#) <https://doi.org/10.1080/01621459.2012.695640>
- LINDQUIST, M. A. and SOBEL, M. E. (2011). Graphical models, potential outcomes and causal inference: Comment on Ramsey, Spirtes and Glymour. *NeuroImage* **57** 334–336.
- LINDSTRÖM, J. (1999). Early development and fitness in birds and mammals. *Trends Ecol. Evol.* **14** 343–348.

- MACKINNON, D. (2012). *Introduction to Statistical Mediation Analysis*. Routledge, London.
- MARMOT, M. G., STANSFELD, S., PATEL, C., NORTH, F., HEAD, J., WHITE, I., BRUNNER, E., FEENEY, A. and SMITH, G. D. (1991). Health inequalities among British civil servants: The Whitehall II study. *Lancet* **337** 1387–1393.
- MCEWEN, B. S. (1998). Stress, adaptation, and disease. Allostasis and allostatic load. *Ann. N.Y. Acad. Sci.* **840** 33–44. <https://doi.org/10.1111/j.1749-6632.1998.tb09546.x>
- MCEWEN, B. S. (2008). Central effects of stress hormones in health and disease: Understanding the protective and damaging effects of stress and stress mediators. *Eur. J. Clin. Pharmacol.* **583** 174–185.
- MILLER, G. E., CHEN, E. and PARKER, K. J. (2011). Psychological stress in childhood and susceptibility to the chronic diseases of aging: Moving toward a model of behavioral and biological mechanisms. *Psychol. Bull.* **137** 959.
- MILLER, G. E., COHEN, S. and RITCHEY, A. K. (2002). Chronic psychological stress and the regulation of pro-inflammatory cytokines: A glucocorticoid-resistance model. *Health Psychology* **21** 531–541. <https://doi.org/10.1037//0278-6133.21.6.531>
- MONTAGNA, S., TOKDAR, S. T., NEELON, B. and DUNSON, D. B. (2012). Bayesian latent factor regression for functional and longitudinal data. *Biometrics* **68** 1064–1073. [MR3040013 https://doi.org/10.1111/j.1541-0420.2012.01788.x](https://doi.org/10.1111/j.1541-0420.2012.01788.x)
- NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (translated). *Statist. Sci.* 465–480.
- NGUYEN, T. Q., SCHMID, I. and STUART, E. A. (2020). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychol. Methods* in press.
- PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* 411–420. Morgan Kaufmann, San Mateo.
- RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR2168993](https://doi.org/10.1007/978-1-4939-9738-9)
- ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 143–155.
- ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**.
- ROSENBAUM, S., ZENG, S., CAMPOS, F., GESQUIERE, L., ALTMANN, J., ALBERTS, S., LI, F. and ARCHIE, E. (2020). Social bonds do not mediate the relationship between early adversity and adult glucocorticoids in wild baboons. *Proc. Natl. Acad. Sci. USA* in press.
- ROTH, D. L. and MACKINNON, D. P. (2012). Mediation analysis with longitudinal data. In *Longitudinal Data Analysis: A Practical Guide for Researchers in Aging, Health, and Social Sciences* 181–216.
- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.
- RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.
- SHPITSER, I. and VANDERWEELE, T. J. (2011). A complete graphical criterion for the adjustment formula in mediation analysis. *Int. J. Biostat.* **7** 16. [MR2775083 https://doi.org/10.2202/1557-4679.1297](https://doi.org/10.2202/1557-4679.1297)
- SILK, J. B. (2007). The adaptive value of sociality in mammalian groups. *Philos. Trans. R. Soc. B, Biol. Sci.* **362** 539–559.
- SILK, J. B., ALTMANN, J. and ALBERTS, S. C. (2006). Social relationships among adult female baboons (*Papio cynocephalus*) I. Variation in the strength of social bonds. *Behav. Ecol. Sociobiol.* **61** 183–195.
- SOBEL, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *J. Educ. Behav. Stat.* **33** 230–251.
- TCHETGEN TCHETGEN, E. J. and SHPITSER, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *Ann. Statist.* **40** 1816–1845. [MR3015045 https://doi.org/10.1214/12-AOS990](https://doi.org/10.1214/12-AOS990)
- TEN HAVE, T. R. and JOFFE, M. M. (2012). A review of causal estimation of effects in mediation analyses. *Stat. Methods Med. Res.* **21** 77–107. [MR2867539 https://doi.org/10.1177/0962280210391076](https://doi.org/10.1177/0962280210391076)
- THORSTEINSSON, E. B. and JAMES, J. E. (1999). A meta-analysis of the effects of experimental manipulations of social support during laboratory stress. *Psychol. Health* **14** 869–886.
- TUNG, J., ARCHIE, E. A., ALTMANN, J. and ALBERTS, S. C. (2016). Cumulative early life adversity predicts longevity in wild baboons. *Nat. Commun.* **7** 1–7.
- VANDERWEELE, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology* **22** 582.
- VANDERWEELE, T. J. (2014). A unification of mediation and interaction: A four-way decomposition. *Epidemiology* **25** 749.
- VANDERWEELE, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford Univ. Press, London.

- VANDERWEELE, T. J. (2016). Mediation analysis: A practitioner's guide. *Annu. Rev. Public Health* **37** 17–32. <https://doi.org/10.1146/annurev-publhealth-032315-021402>
- VANDERWEELE, T. J. and TCHETGEN TCHETGEN, E. J. (2017). Mediation analysis with time varying exposures and mediators. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 917–938. MR3641414 <https://doi.org/10.1111/rssb.12194>
- VANDERWEELE, T. J., VANSTEEELANDT, S. and ROBINS, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology* **25** 300.
- VANSTEEELANDT, S., LINDER, M., VANDENBERGHE, S., STEEN, J. and MADSEN, J. (2019). Mediation analysis of time-to-event endpoints accounting for repeatedly measured mediators subject to time-varying confounding. *Stat. Med.* **38** 4828–4840. MR4022831 <https://doi.org/10.1002/sim.8336>
- VAN DER LAAN, M. J. and PETERSEN, M. L. (2008). Direct effect models. *Int. J. Biostat.* **4** 23. MR2456975 <https://doi.org/10.2202/1557-4679.1064>
- YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561 <https://doi.org/10.1198/016214504000001745>
- ZENG, S., ROSENBAUM, S., ARCHIE, E., ALBERTS, S. and LI, F. (2021). Supplement to “Causal mediation analysis for sparse and irregular longitudinal data.” <https://doi.org/10.1214/20-AOAS1427SUPPA>, <https://doi.org/10.1214/20-AOAS1427SUPPB>
- ZHAO, Y., LUO, X., LINDQUIST, M. and CAFFO, B. (2018). Functional mediation analysis with an application to functional magnetic resonance imaging data. *Preprint. Available at arXiv:1805.06923*.
- ZHENG, W. and VAN DER LAAN, M. (2017). Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes. *Journal of Causal Inference* **5**.
- ZIGLER, C. M., DOMINICI, F. and WANG, Y. (2012). Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics* **13** 289–302.
- ZIPPLE, M. N., ARCHIE, E. A., TUNG, J., ALTMANN, J. and ALBERTS, S. C. (2019). Intergenerational effects of early adversity on survival in wild baboons. *eLife* **8** e47433.

RAPID DESIGN OF METAMATERIALS VIA MULTITARGET BAYESIAN OPTIMIZATION

BY YANG YANG¹, CHUNLIN JI² AND KE DENG³

¹Department of Mathematical Sciences & Center for Statistical Science, Tsinghua University, yyang15@mails.tsinghua.edu.cn

²Kuang-Chi Institute of Advanced Technology, Shenzhen, China, chunlin.ji@kuang-chi.com

³Center for Statistical Science and Department of Industrial Engineering, Tsinghua University, kdeng@tsinghua.edu.cn

Composed of a large number of subwavelength unit cells with designable geometries, metamaterials have been widely studied to achieve extraordinary advantageous and unusual optical properties. However, ordinary computer simulator requires a time-consuming fine-tuning to find a proper design of metamaterial for a specific optical property, making the design stage a critical bottleneck in large scale applications of metamaterials. This paper investigates the metamaterial design under the framework of computer experiments, with emphasis on dealing with the challenge of designing numerous unit cells with functional responses, simultaneously, which is not common in traditional computer experiments. We formulate the multiple related design targets as a multitarget design problem. Leveraging on the similarity between different designs, we propose an efficient Bayesian optimization strategy with a parsimonious surrogate model and an integrated acquisition function to design multiple unit cells with very few function evaluations. A wide range of simulations confirm the effectiveness and superiority of the proposed approach compared to the naive strategies where the multiple unit cells are dealt with separately or sequentially. Such a rapid design strategy has the potential to greatly promote large scale applications of metamaterials in practice.

REFERENCES

- BALI, K. K., ONG, Y., GUPTA, A. and TAN, P. S. (2020). Multifactorial evolutionary algorithm with online transfer parameter estimation: MFEA-II. *IEEE Trans. Evol. Comput.* **24** 69–83.
- BILOTTI, F., TOSCANO, A. and VEGNI, L. (2007). Design of spiral and multiple split-ring resonators for the realization of miniaturized metamaterial samples. *IEEE Trans. Antennas and Propagation* **55** 2258–2267.
- BOYLE, P. and FREAN, M. (2004). Multiple output Gaussian process regression. Technical Report.
- BULL, A. D. (2011). Convergence rates of efficient global optimization algorithms. *J. Mach. Learn. Res.* **12** 2879–2904. [MR2854351](#)
- CHEN, H., RAN, L., HUANGFU, J., GRZEGORCZYK, T. M. and KONG, J. A. (2006). Equivalent circuit model for left-handed metamaterials. *J. Appl. Phys.* **100** 024915.
- CUI, T., SMITH, D. R. and LIU, R. (2010). *Metamaterials: Theory, Design, and Applications*. Springer, New York.
- DA, B., ONG, Y. S., FENG, L., QIN, A. K., GUPTA, A., ZHU, Z., TING, C. K., TANG, K. and YAO, X. (2016). Evolutionary multitasking for single-objective continuous optimization: Benchmark problems, performance metrics and baseline results. Technical Report.
- DEB, K., PRATAP, A., AGARWAL, S. and MEYARIVAN, T. (2002). A fast and elitist multi-objective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **6** 182–197.
- FANG, K.-T., LIN, D. K. J., WINKER, P. and ZHANG, Y. (2000). Uniform design: Theory and application. *Technometrics* **42** 237–248. [MR1801031](#) <https://doi.org/10.2307/1271079>
- FRAZIER, P. I. and WANG, J. (2016). Bayesian optimization for materials design. In *Information Science for Materials Discovery and Design* 45–75. Springer, Berlin.
- GINSBOURGER, D., RICHE, R. L. and CARRARO, L. (2008). A multi-points criterion for deterministic parallel global optimization based on Gaussian processes. HAL Preprint Hal-00260579.

- GUO, X., HE, Y., ZHU, B., YANG, Y., DENG, K., LIU, R. and JI, C. (2017a). Bayesian uncertainty quantification for functional response. In *Uncertainty Quantification and Model Calibration* **6** 115–131. IntechOpen.
- GUO, X., JI, C., LIU, R. and TANG, T. (2017b). A two-stage approach for frequency response modeling and metamaterial rapid design. *Prog. Electromagn. Res. C* **76** 11–22.
- GUPTA, A., ONG, Y. S. and FENG, L. (2016). Multifactorial evolution: Toward evolutionary multitasking. *IEEE Trans. Evol. Comput.* **20** 343–357.
- HAUPT, R. L. and HAUPT, S. E. (2004). *Practical Genetic Algorithms*, 2nd ed. Wiley Interscience, Hoboken, NJ. [MR2059939](#)
- HENNIG, P. and SCHULER, C. J. (2012). Entropy search for information-efficient global optimization. *J. Mach. Learn. Res.* **13** 1809–1837. [MR2956343](#)
- HIGDON, D., KENNEDY, M., CAVENDISH, J. C., CAFFEO, J. A. and RYNE, R. D. (2004). Combining field data and computer simulations for calibration and prediction. *SIAM J. Sci. Comput.* **26** 448–466. [MR2116355](#) <https://doi.org/10.1137/S1064827503426693>
- HWANG, C. L. and MASUD, A. S. M. (1979). *Multiple Objective Decision Making—Methods and Applications: A State-of-the-Art Survey. Lecture Notes in Economics and Mathematical Systems* **164**. Springer, Berlin. [MR0567888](#)
- JONES, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *J. Global Optim.* **21** 345–383. [MR1869398](#) <https://doi.org/10.1023/A:1012771025575>
- JONES, D. R., SCHONLAU, M. and WELCH, W. J. (1998). Efficient global optimization of expensive black-box functions. *J. Global Optim.* **13** 455–492. [MR1673460](#) <https://doi.org/10.1023/A:1008306431147>
- JU, S., SHIGA, T., FENG, L., HOU, Z., TSUDA, K. and SHIOMI, J. (2017). Designing nanostructures for phonon transport via Bayesian optimization. *Phys. Rev. X* **7** 021024.
- KENNEDY, M. C. and O'HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. [MR1858398](#) <https://doi.org/10.1111/1467-9868.00294>
- KIYOHARA, S., ODA, H., TSUDA, K. and MIZOGUCHI, T. (2016). Acceleration of stable interface structure searching using a kriging approach. *Jpn. J. Appl. Phys.* **55** 045502.
- KLEIJNEN, J. P. C. (2009). Kriging metamodeling in simulation: A review. *European J. Oper. Res.* **192** 707–716. [MR2457613](#) <https://doi.org/10.1016/j.ejor.2007.10.013>
- KONAK, A., COIT, D. W. and SMITH, A. E. (2006). Multi-objective optimization using genetic algorithms: A tutorial. *Reliab. Eng. Syst. Saf.* **91** 992–1007.
- LI, L. W., LI, Y. N., YEO, T. S., MOSIG, J. R. and MARTIN, O. J. F. (2010). A broadband and high-gain metamaterial microstrip antenna. *Appl. Phys. Lett.* **96** 164101.
- LI, D., ZIGONEANU, L., POPA, B. I. and CUMMER, S. A. (2012). Design of an acoustic metamaterial lens using genetic algorithms. *J. Acoust. Soc. Am.* **132** 2823–2833.
- LIU, B. and JI, C. (2014). Bayesian nonparametric modeling for rapid design of metamaterial microstructures. *Int. J. Antennas Propag.* **2014** 165102.
- LIU, R., JI, C., MOCK, J. J., CHIN, J. Y., CUI, T. J. and SMITH, D. R. (2009). Broadband ground-plane cloak. *Science* **323** 366–369.
- MAJUMDAR, P., ZHAO, Z., YUE, Y., JI, C. and LIU, R. (2014). Equivalent circuit model of different configurations of loop elements using vector-fitting. In *PIERS Proceedings* 25–28.
- MARLER, R. T. and ARORA, J. S. (2004). Survey of multi-objective optimization methods for engineering. *Struct. Multidiscip. Optim.* **26** 369–395. [MR2057377](#) <https://doi.org/10.1007/s00158-003-0368-6>
- MEBANE, W. R. and SEKHON, J. S. (2011). Genetic optimization using derivatives: The rgenoud package for R. *J. Stat. Softw.* **42** 1–26.
- MOKKUS, J. (1975). On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference* 400–404.
- MURATA, T. and ISHIBUCHI, H. (1995). MOGA: Multi-objective genetic algorithms. In *IEEE International Conference on Evolutionary Computation* **1**.
- OLSSON, D. M. and NELSON, L. S. (1975). The Nelder–Mead simplex procedure for function minimization. *Technometrics* **17** 45–51.
- OWEN, A. B. (1992). Orthogonal arrays for computer experiments, integration and visualization. *Statist. Sinica* **2** 439–452. [MR1187952](#)
- PACKWOOD, D. (2017). *Bayesian Optimization for Materials Science. SpringerBriefs in the Mathematics of Materials* **3**. Springer, Singapore. [MR3701411](#) <https://doi.org/10.1007/978-981-10-6781-5>
- PACKWOOD, D. M. and HITOSUGI, T. (2017). Rapid prediction of molecule arrangements on metal surfaces via Bayesian optimization. *Appl. Phys. Express* **10** 065502.
- PADILLA, W. J., BASOV, D. N. and SMITH, D. R. (2006). Negative refractive index metamaterials. *Mater. Today* **9** 28–35.
- PENDRY, J. B., SCHURIG, D. and SMITH, D. R. (2006). Controlling electromagnetic fields. *Science* **312** 1780–1782. [MR2237570](#) <https://doi.org/10.1126/science.1125907>

- ROUSTANT, O., GINSBOURGER, D. and DEVILLE, Y. (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *J. Stat. Softw.* **51** 1–55.
- SANDERS, G. H. W. and MANZ, A. (2000). Chip-based microsystems for genomic and proteomic analysis. *TrAC, Trends Anal. Chem.* **19** 364–378.
- SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2003). *The Design and Analysis of Computer Experiments. Springer Series in Statistics*. Springer, New York. MR2160708 <https://doi.org/10.1007/978-1-4757-3799-8>
- SCHAFER, J. D. (1987). Multiple objective optimization with vector evaluated genetic algorithms. In *Proceedings of the First International Conference on Genetic Algorithms* 93–100.
- SCOTT, W., FRAZIER, P. and POWELL, W. (2011). The correlated knowledge gradient for simulation optimization of continuous parameters using Gaussian process regression. *SIAM J. Optim.* **21** 996–1026. MR2837561 <https://doi.org/10.1137/100801275>
- SEKO, A., TOGO, A., HAYASHI, H., TSUDA, K., CHAPUT, L. and TANAKA, I. (2015). Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization. *Phys. Rev. Lett.* **115** 205901.
- SHAHRIARI, B., SWERSKY, K., WANG, Z. Y., ADAMS, R. R. and DE FREITAS, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* **104** 148–175.
- SHALAEV, V. M. (2007). Optical negative-index metamaterials. *Nat. Photonics* **1** 41–48.
- SINGH, G. and MARWAHA, A. (2015). A review of metamaterials and its applications. *Int. J. Eng. Trends Technol.* **19** 305–310.
- SRINIVAS, N., KRAUSE, A., KAKADE, S. and SEEGER, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning* 1015–1022.
- SUN, F., WANG, Y. and XU, H. (2019). Uniform projection designs. *Ann. Statist.* **47** 641–661. MR3909945 <https://doi.org/10.1214/18-AOS1705>
- SWERSKY, K., SNOEK, J. and ADAMS, R. P. (2013). Multi-task Bayesian optimization. In *Advances in Neural Information Processing Systems* 2004–2012.
- TANG, B. (1993). Orthogonal array-based Latin hypercubes. *J. Amer. Statist. Assoc.* **88** 1392–1397. MR1245375
- TAO, H., LANDY, N. I., BINGHAM, C. M., ZHANG, X., AVERITT, R. D. and PADILLA, W. J. (2008). A metamaterial absorber for the terahertz regime: Design, fabrication and characterization. *Opt. Express* **16** 7181–7188.
- UENO, T., RHONE, T. D., HOU, Z., MIZOGUCHI, T. and TSUDA, K. (2016). COMBO: An efficient Bayesian optimization library for materials science. *Mater. Discov.* **4** 18–21.
- VALENTINE, J., ZHANG, S., ZENTGRAF, T., ULIN-AVILA, E., GENOV, D. A., BARTAL, G. and ZHANG, X. (2008). Three-dimensional optical metamaterial with a negative refractive index. *Nature* **455** 376–379.
- VAZQUEZ, E. and BECT, J. (2010). Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *J. Statist. Plann. Inference* **140** 3088–3095. MR2659839 <https://doi.org/10.1016/j.jspi.2010.04.018>
- XIONG, S., QIAN, P. Z. G. and WU, C. F. J. (2013). Sequential design and analysis of high-accuracy and low-accuracy computer codes. *Technometrics* **55** 37–46. MR3038483 <https://doi.org/10.1080/00401706.2012.723572>
- YANG, Y., JI, C. and DENG, K. (2021). Supplement to “Rapid design of metamaterials via multitarget Bayesian optimization.” <https://doi.org/10.1214/20-AOAS1426SUPP>
- YAO, Y., ALLARDYCE, B. J., RAJKHOWA, R., HEGH, D., SUTTI, A., SUBIANTO, S., GUPTA, S., RANA, S., GREENHILL, S. et al. (2020). Improving the tensile properties of wet spun silk fibers using rapid Bayesian algorithm. *ACS Biomater. Sci. Eng.* **6** 3197–3207.
- YE, K. Q. (1998). Orthogonal column Latin hypercubes and their application in computer experiments. *J. Amer. Statist. Assoc.* **93** 1430–1439. MR1666638 <https://doi.org/10.2307/2670057>
- ZHAO, Y., CHEN, F., CHEN, H., LI, N., SHEN, Q. and ZHANG, L. (2011). The microstructure design optimization of negative index metamaterials using genetic algorithm. *PIER Lett.* **22** 95–108.
- ZHENG, X., QIN, A. K., GONG, M. and ZHOU, D. (2019). Self-regulated evolutionary multi-task optimization. *IEEE Trans. Evol. Comput.* **24** 16–28.

A CONTINUOUS-TIME SEMI-MARKOV MODEL FOR ANIMAL MOVEMENT IN A DYNAMIC ENVIRONMENT

BY DEVIN JOHNSON^{*}, NOEL PELLAND[†] AND JEREMY STERLING[‡]

Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, ^{*}devin.johnson@noaa.gov;

[†]noel.pelland@noaa.gov; [‡]jeremy.sterling@noaa.gov

We consider an extension to discrete-space, continuous-time models for animal movement that have previously been presented in the literature. The extension from a continuous-time Markov formulation to a continuous-time *semi*-Markov formulation allows for the inclusion of temporally dynamic habitat conditions as well as temporally changing movement responses by animals to that environment. We show that, with only a little additional consideration, the Poisson likelihood calculation for the Markov version can still be used within the multiple imputation framework commonly employed for analysis of telemetry data. In addition, we consider a Bayesian model selection methodology within the imputation framework. The model selection method uses a Laplace approximation to the posterior model probability to provide a computationally feasible approach. The full methodology is then used to analyze movements of 15 weaned northern fur seal (*Callorhinus ursinus*) pups with respect to surface winds, geostrophic currents and sea surface temperature. The highest posterior model probabilities belonged to those models containing only winds and current; SST was not a significant factor for modeling their movement.

REFERENCES

- AVGAR, T., POTTS, J. R., LEWIS, M. A. and BOYCE, M. S. (2016). Integrated step selection analysis: Bridging the gap between resource selection and animal movement. *Methods Ecol. Evol.* **7** 619–630.
- BAYARRI, M. J., BERGER, J. O., FORTE, A. and GARCÍA-DONATO, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Ann. Statist.* **40** 1550–1577. [MR3015035](#) <https://doi.org/10.1214/12-AOS1013>
- BENGTTSSON, H. (2017). doFuture: A universal foreach parallel adaptor using the future API of the ‘future’ package. R package version 0.6.0.
- BENGTTSSON, H. (2018). future: Unified parallel and distributed processing in R for everyone. R package version 1.8.1.
- BOOTH, C. G., SINCLAIR, R. R. and HARWOOD, J. (2020). Methods for monitoring for the population consequences of disturbance in marine mammals: A review. *Front. Mar. Sci.* **7** 115.
- BUDERMAN, F. E., HOOTEN, M. B., ALLDREDGE, M., HANKS, E. M. and IVAN, J. S. (2018). Predatory behavior is primary predictor of movement of wildland-urban cougars. *bioRxiv* 257295.
- COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data* **21**. CRC Press.
- HANKS, E. M., HOOTEN, M. B. and ALLDREDGE, M. W. (2015). Continuous-time discrete-space models for animal movement. *Ann. Appl. Stat.* **9** 145–165. [MR3341111](#) <https://doi.org/10.1214/14-AOAS803>
- HANKS, E. M. and HUGHES, D. A. (2016). Flexible discrete space models of animal movement. [arXiv:1606.07986](#).
- HOLFORD, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics* **36** 299–305.
- HOOPER, C. L. (1895). Report of Captain Hooper, dated November 21, 1892. In *US Senate, 53rd Congress, 2nd Session. Fur-seal Arbitration, Proceedings of the Tribunal of Arbitration, Volume VII, Senate Executive Document 177 (serial Set 3166)* 228–233. Government Printing Office, Washington, D.C.
- HOOTEN, M. B., JOHNSON, D. S. and BROST, B. M. (2019). Making recursive Bayesian inference accessible. *Amer. Statist.* <https://doi.org/10.1080/00031305.2019.1665584>

- HOOTEN, M. B., JOHNSON, D. S., HANKS, E. M. and LOWRY, J. H. (2010). Agent-based inference for animal movement and selection. *J. Agric. Biol. Environ. Stat.* **15** 523–538. [MR2788638](#) <https://doi.org/10.1007/s13253-010-0038-2>
- HOOTEN, M. B., BUDERMAN, F. E., BROST, B. M., HANKS, E. M. and IVAN, J. S. (2016). Hierarchical animal movement models for population-level inference. *Environmetrics* **27** 322–333. [MR3545640](#) <https://doi.org/10.1002/env.2402>
- HOOTEN, M. B., JOHNSON, D. S., MCCLINTOCK, B. T. and MORALES, J. M. (2017). *Animal Movement: Statistical Models for Telemetry Data*. Chapman and Hall/CRC, Boca Raton, FL.
- HOOVER, A. L., LIANG, D., ALFARO-SHIGUETO, J., MANGEL, J. C., MILLER, P. I., MORREALE, S. J., BAILEY, H. and SHILLINGER, G. L. (2019). Predicting residence time using a continuous-time discrete-space model of leatherback turtle satellite telemetry data. *Ecosphere* **10** e02644.
- 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. and LONDON, J. M. (2018). *crawl*: An R package for fitting continuous-time correlated random walk models to animal movement data. <https://doi.org/10.5281/zenodo.596464>
- JOHNSON, D. S., LONDON, J. M. and KUHN, C. E. (2011). Bayesian inference for animal space use and other movement metrics. *J. Agric. Biol. Environ. Stat.* **16** 357–370. [MR2843131](#) <https://doi.org/10.1007/s13253-011-0056-8>
- JOHNSON, D., PELLAND, N. and STERLING, J. (2021). Supplement to “A continuous-time semi-markov model for animal movement in a dynamic environment.” <https://doi.org/10.1214/20-AOAS1408SUPPA>, <https://doi.org/10.1214/20-AOAS1408SUPPB>, <https://doi.org/10.1214/20-AOAS1408SUPPC>
- JOHNSON, D. S., LONDON, J. M., LEA, M. A. and DURBAN, J. W. (2008). Continuous-time correlated random walk model for animal telemetry data. *Ecology* **89** 1208–1215.
- KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. [MR3363402](#) <https://doi.org/10.1080/01621459.1995.10476572>
- LAIRD, N. and OLIVIER, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *J. Amer. Statist. Assoc.* **76** 231–240. [MR0624329](#)
- LEA, M. A., JOHNSON, D., REAM, R., STERLING, J., MELIN, S. and GELATT, T. (2009). Extreme weather events influence dispersal of naive northern fur seals. *Biol. Lett.* **5** 252–257.
- LIU, Y., BATTAILLE, B. C., ZIDEK, J. V. and TRITES, A. (2014). Bayesian melding of the dead-reckoned path and GPS measurements for an accurate and high-resolution path of marine mammals. [arXiv:1411.6683](#).
- MCCRACKEN, M. L., MANLY, B. F. J. and VANDER HEYDEN, M. (1998). The use of discrete-choice models for evaluating resource selection. *J. Agric. Biol. Environ. Stat.* **3** 268–279. [MR1817038](#) <https://doi.org/10.2307/1400582>
- MICROSOFT and WESTON, S. (2017). *foreach*: Provides foreach looping construct for R. R package version 1.4.4.
- MILLER, D. L. (2019). Bayesian views of generalized additive modelling. [arXiv:1902.01330](#).
- NAKAGAWA, S. and FRECKLETON, R. P. (2011). Model averaging, missing data and multiple imputation: A case study for behavioural ecology. *Behav. Ecol. Sociobiol.* **65** 103–116.
- NORRIS, J. R. (1998). *Markov Chains. Cambridge Series in Statistical and Probabilistic Mathematics* **2**. Cambridge Univ. Press, Cambridge. Reprint of 1997 original. [MR1600720](#)
- PELLAND, N. A., STERLING, J. T., LEA, M. A., BOND, N. A., REAM, R. R., LEE, C. M. and ERIKSEN, C. C. (2014). Fortuitous encounters between seagliders and adult female northern fur seals (*Callorhinus ursinus*) off the Washington (USA) coast: Upper ocean variability and links to top predator behavior. *PLoS ONE* **9** e101268.
- R CORE TEAM (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- REAM, R. R., STERLING, J. T. and LOUGHLIN, T. R. (2005). Oceanographic features related to northern fur seal migratory movements. *Deep-Sea Res., Part 2, Top. Stud. Oceanogr.* **52** 823–843.
- RUBIN, D. B. and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Amer. Statist. Assoc.* **81** 366–374. [MR0845877](#)
- SCHARF, H., HOOTEN, M. B. and JOHNSON, D. S. (2017). Imputation approaches for animal movement modeling. *J. Agric. Biol. Environ. Stat.* **22** 335–352. [MR3692468](#) <https://doi.org/10.1007/s13253-017-0294-5>
- STERLING, J. T., SPRING, A. M., IVERSON, S. J., JOHNSON, S. P., PELLAND, N. A., JOHNSON, D. S., LEA, M. A. and BOND, N. A. (2014). The sun, moon, wind, and biological imperative—shaping contrasting wintertime migration and foraging strategies of adult male and female northern fur seals (*Callorhinus ursinus*). *PLoS ONE* **9** e93068.
- VALLIS, G. K. (2006). *Atmospheric and Oceanic Fluid Dynamics*. Cambridge Univ. Press.
- VOLINSKY, C. T., MADIGAN, D., RAFTERY, A. E. and KRONMAL, R. A. (1997). Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **46** 433–448.

- WHITE, I. R., ROYSTON, P. and WOOD, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **30** 377–399. MR2758870 <https://doi.org/10.1002/sim.4067>
- WOOD, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 3–36. MR2797734 <https://doi.org/10.1111/j.1467-9868.2010.00749.x>

A BAYESIAN SEMIPARAMETRIC JOLLY–SEBER MODEL WITH INDIVIDUAL HETEROGENEITY: AN APPLICATION TO MIGRATORY MALLARDS AT STOPOVER

BY GUOHUI WU¹, SCOTT H. HOLAN², ALEXIS AVRIL^{3,*} AND JONAS WALDENSTRÖM^{3,†}

¹SAS Institute Inc., raywu2014@gmail.com

²Department of Statistics, University of Missouri, holans@missouri.edu

³Centre for Ecology and Evolution in Microbial Model Systems, Linnaeus University, *alexis.avril@gmail.com;

†jonas.waldenstrom@lnu.se

We propose a Bayesian hierarchical Jolly–Seber model that can accommodate a semiparametric functional relationship between external covariates and capture probabilities, individual heterogeneity in departure due to an internal time-varying covariate and the dependence of arrival time on external covariates. Modelwise, we consider a stochastic process to characterize the evolution of the partially observable internal covariate that is linked to departure probabilities. Computationally, we develop a well-tailored Markov chain Monte Carlo algorithm that is free of tuning through data augmentation. Inferentially, our model allows us to make inference about stopover duration and population sizes, the impacts of various covariates on departure and arrival time and to identify flexible yet data-driven functional relationships between external covariates and capture probabilities. We demonstrate the effectiveness of our model through a motivating dataset collected for studying the migration of mallards (*Anas platyrhynchos*) in Sweden.

REFERENCES

- BENGTSSON, D., AVRIL, A., GUNNARSSON, G., ELMBERG, J., SÖDERQUIST, P., NOREVIK, G., TOLF, C., SAFI, K., FIEDLER, W. et al. (2014). Movements, home-range size and habitat selection of mallards during autumn migration. *PLoS ONE* **9** e100764. <https://doi.org/10.1371/journal.pone.0100764>
- BERTHOLD, P. (2001). *Bird Migration: A General Survey*, 2nd ed. Oxford Univ. Press, New York.
- BONNER, S. J. and SCHWARZ, C. J. (2006). An extension of the Cormack–Jolly–Seber model for continuous covariates with application to *Microtus pennsylvanicus*. *Biometrics* **62** 142–149. MR2226567 <https://doi.org/10.1111/j.1541-0420.2005.00399.x>
- BONNER, S. J. and SCHWARZ, C. J. (2009). Bayesian estimation of abundance for open populations with covariate dependent capture and survival probabilities draft. Available at http://www.simon.bonnors.ca/PDFS/cts_covariate_js_0809.pdf.
- BONNER, S. J. and SCHWARZ, C. J. (2011). Smoothing population size estimates for time-stratified mark-recapture experiments using Bayesian P-splines. *Biometrics* **67** 1498–1507. MR2872401 <https://doi.org/10.1111/j.1541-0420.2011.01599.x>
- BONNER, S., THOMSON, D. and SCHWARZ, C. (2009). Time-varying covariates and semi-parametric regression in capture–recapture: An adaptive spline approach. In *Modeling Demographic Processes in Marked Populations, Environmental and Ecological Statistics* 3 (D. Thomson, E. Cooch and M. Conroy, eds.) 657–675. Springer, New York.
- BROOKS, S. P., CATCHPOLE, E. A. and MORGAN, B. J. T. (2000). Bayesian animal survival estimation. *Statist. Sci.* **15** 357–376. MR1847773 <https://doi.org/10.1214/ss/1009213003>
- BROOKS, S. P. and GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7** 434–455. MR1665662 <https://doi.org/10.2307/1390675>
- CRAINICEANU, C. M., RUPPERT, D. and WAND, M. P. (2005). Bayesian analysis for penalized spline regression using Win BUGS. *J. Stat. Softw.* **14** 1–24.

Key words and phrases. Capture–recapture, individual heterogeneity, low-rank thin-plate splines, Ornstein–Uhlenbeck process, stopover duration analysis.

- CRAMP, S. and SIMMONS, K. (1977). *The Birds of the Western Palearctic. Handbook of the Birds of Europe, the Middle East and North Africa. Ostrich to Ducks* **1**. Oxford University Press, Oxford.
- DINSMORE, S. J. and COLLAZO, J. A. (2003). The influence of body condition on local apparent survival of spring migrant sanderlings in coastal North Carolina. *Condor* **105** 465–473.
- FILIPPE, P. A., BRAUMANN, C. A., BRITES, N. M. and ROQUETE, C. J. (2010). Modelling animal growth in random environments: An application using nonparametric estimation. *Biom. J.* **52** 653–666. [MR2757011 https://doi.org/10.1002/bimj.200900273](https://doi.org/10.1002/bimj.200900273)
- FINCH, S. (2004). Ornstein–Uhlenbeck process. Available at <https://oeis.org/A249417/a249417.pdf>.
- GELMAN, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *Int. Stat. Rev.* **71** 369–382.
- GIMENEZ, O., CRAINICEANU, C., BARBRAUD, C., JENOUVRIER, S. and MORGAN, B. J. T. (2006). Semi-parametric regression in capture–recapture modeling. *Biometrics* **62** 691–698. [MR2247195 https://doi.org/10.1111/j.1541-0420.2005.00514.x](https://doi.org/10.1111/j.1541-0420.2005.00514.x)
- GUNNARSSON, G., LATORRE-MARGALEF, N., HOBSON, K. A., WILGENBURG, S. L. V., ELMBERG, J., OLSEN, B., FOUCHIER, R. A. M. and WALDENSTRÖM, J. (2012). Disease dynamics and bird migration—linking mallards *Anas platyrhynchos* and subtype diversity of the influenza A virus in time and space. *PLoS ONE* **7** e35679. <https://doi.org/10.1371/journal.pone.0035679>
- JOLLY, G. M. (1965). Explicit estimates from capture–recapture data with both death and immigration-stochastic model. *Biometrika* **52** 225–247. [MR0210227 https://doi.org/10.1093/biomet/52.1-2.225](https://doi.org/10.1093/biomet/52.1-2.225)
- KAISER, A. (1995). Estimating turnover, movements and capture parameters of resting passerines in standardized capture–recapture studies. *J. Appl. Stat.* **22** 1039–1048.
- KING, R. and BROOKS, S. P. (2002). Bayesian model discrimination for multiple strata capture–recapture data. *Biometrika* **89** 785–806. [MR1946510 https://doi.org/10.1093/biomet/89.4.785](https://doi.org/10.1093/biomet/89.4.785)
- KING, R., MORGAN, B., GIMENEZ, O. and BROOKS, S. (2010). *Bayesian Analysis for Population Ecology*. CRC Press/CRC, Boca Raton, FL.
- KUENZI, A. J., MOORE, F. R. and SIMONS, T. R. (1991). Stopover of Neotropical landbird migrants on East Ship Island following trans-Gulf migration. *Condor* **93** 869–883.
- LATORRE-MARGALEF, N., GUNNARSSON, G., MUNSTER, V. J., FOUCHIER, R. A., OSTERHAUS, A. D., ELMBERG, J., OLSEN, B., WALLENSTEN, A., HAEMIG, P. D. et al. (2009). Effects of influenza A virus infection on migrating mallard ducks. *Proc. R. Soc. Lond., B Biol. Sci.* **276** 1029–1036.
- LUNN, D. J., THOMAS, A., BEST, N. and SPIEGELHALTER, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat. Comput.* **10** 325–337.
- LYONS, J. E., KENDALL, W. L., ROYLE, J. A., CONVERSE, S. J., ANDRES, B. A. and BUCHANAN, J. B. (2016). Population size and stopover duration estimation using mark-resight data and Bayesian analysis of a superpopulation model. *Biometrics* **72** 262–271. [MR3500595 https://doi.org/10.1111/biom.12393](https://doi.org/10.1111/biom.12393)
- MATECHOU, E. (2010). Applications and extensions of capture–recapture stop-over models. Ph.D. thesis, The University of Kent.
- MATECHOU, E., DENNIS, E. B., FREEMAN, S. N. and BRERETON, T. (2014). Monitoring abundance and phenology in (multivoltine) butterfly species: A novel mixture model. *J. Appl. Ecol.* **51** 766–775.
- MATECHOU, E., NICHOLLS, G. K., MORGAN, B. J. T., COLLAZO, J. A. and LYONS, J. E. (2016). Bayesian analysis of Jolly–Seber type models. *Environ. Ecol. Stat.* **23** 531–547. [MR3568110 https://doi.org/10.1007/s10651-016-0352-0](https://doi.org/10.1007/s10651-016-0352-0)
- NEWTON, I. (2010). *The Migration Ecology of Birds*. Academic Press, London.
- NICHOLS, J. D., RUNGE, M. C., JOHNSON, F. A. and WILLIAMS, B. K. (2007). Adaptive harvest management of North American waterfowl populations: A brief history and future prospects. *J. Ornithol.* **148** 343–349.
- PLEDGER, S., EFFORD, M., POLLOCK, K., COLLAZO, J. and LYONS, J. (2009). Stopover duration analysis with departure probability dependent on unknown time since arrival. In *Environmental and Ecological Statistics* (D. Thomson, E. Cooch and M. Conroy, eds.) 349–363. Springer, Berlin.
- PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6** 7–11.
- RICE, S. M., COLLAZO, J. A., ALLDREDGE, M. W., HARRINGTON, B. A., LEWIS, A. R. and SANDERCOCK, B. (2007). Local annual survival and seasonal residency rates of semipalmated sandpipers (*Calidris pusilla*) in Puerto Rico. *Auk* **124** 1397–1406.
- ROYLE, J. A. and DORAZIO, R. M. (2008). *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press, San Diego, CA.
- RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. [MR1998720 https://doi.org/10.1017/CBO9780511755453](https://doi.org/10.1017/CBO9780511755453)
- SCHAUB, M., JENNI, L. and BAIRLEIN, F. (2008). Fuel stores, fuel accumulation, and the decision to depart from a migration stopover site. *Behav. Ecol.* **19** 657–666.

- SCHOFIELD, M. R. and BARKER, R. J. (2011). Full open population capture–recapture models with individual covariates. *J. Agric. Biol. Environ. Stat.* **16** 253–268. [MR2818548](#) <https://doi.org/10.1007/s13253-010-0052-4>
- 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](#) <https://doi.org/10.2307/2533048>
- SEBER, G. A. F. (1965). A note on the multiple-recapture census. *Biometrika* **52** 249–259. [MR0210228](#) <https://doi.org/10.1093/biomet/52.1-2.249>
- TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. [MR0898357](#)
- R CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- VAN TOOR, M. L., AVRIL, A., WU, G., HOLAN, S. H. and WALDENSTRÖM, J. (2018a). As the duck flies—Estimating the dispersal of low-pathogenic avian influenza viruses by migrating mallards. *Front. Ecol. Evol.* **6** 208.
- VAN TOOR, M. L., KRANSTAUBER, B., NEWMAN, S. H., PROSSER, D. J., TAKEKAWA, J. Y., TECHNITIS, G., WEIBEL, R., WIKELSKI, M. and SAFI, K. (2018b). Integrating animal movement with habitat suitability for estimating dynamic migratory connectivity. *Landsc. Ecol.* **33** 879–893.
- VON BERTALANFFY, L. K. (1938). A quantitative theory of organic growth (inquiries on growth laws. II). *Hum. Biol.* **10** 181–213.
- WILLIAMS, B., NICHOLS, J. and CONROY, M. (2002). *Analysis and Management of Animal Populations*. Academic Press, San Diego, CA.
- WU, G., HOLAN, S. H., AVRIL, A. and WALDENSTRÖM, J. (2021). Supplement to “A Bayesian semiparametric Jolly–Seber model with individual heterogeneity: An application to migratory mallards at stopover.” <https://doi.org/10.1214/20-AOAS1421SUPPA>, <https://doi.org/10.1214/20-AOAS1421SUPPB>
- YONG, W. and MOORE, F. R. (1993). Relation between migratory activity and energetic condition among thrushes (Turdinae) following passage across the Gulf of Mexico. *Condor* **95** 934–943.

INFERENCE OF LARGE MODIFIED POISSON-TYPE GRAPHICAL MODELS: APPLICATION TO RNA-SEQ DATA IN CHILDHOOD ATOPIC ASTHMA STUDIES

BY RONG ZHANG^{1,*}, ZHAO REN^{1,†}, JUAN C. CELEDÓN^{2,‡} AND WEI CHEN^{2,§}

¹*Department of Statistics, University of Pittsburgh, *roz16@pitt.edu; †zren@pitt.edu*

²*Department of Pediatrics, UPMC Children's Hospital of Pittsburgh, University of Pittsburgh, ‡juan.celedon@chp.edu; §wei.chen@chp.edu*

Recent advances in next-generation sequencing technology have yielded huge amounts of transcriptomic data. The discreteness and the high dimensions of RNA-seq data have posed great challenges in biological network analysis. Although estimation theories for high-dimensional modified Poisson-type graphical models have been proposed for the network analysis of count-valued data, the statistical inference of these models is still largely unknown. We herein propose a two-step procedure in both edgewise and global statistical inference of these modified Poisson-type graphical models using a cutting-edge generalized low-dimensional projection approach for bias correction. Extensive simulations and a real example with ground truth illustrate asymptotic normality of edgewise inference and more accurate inferential results in multiple testing compared to the sole estimation and the inferential method under normal assumption. Furthermore, the application of our method to novel RNA-seq data of childhood atopic asthma in Puerto Ricans demonstrates more biologically meaningful results compared to the sole estimation and the inferential methods based on Gaussian and nonparanormal graphical models.

REFERENCES

- ALLEN, G. I. and LIU, Z. (2013). A local Poisson graphical model for inferring networks from sequencing data. *IEEE Trans. Nanobiosci.* **12** 189–198.
- ALMAAS, E. and BARABÁSI, A.-L. (2006). Power laws in biological networks. In *Power Laws, Scale-Free Networks and Genome Biology* 1–11. Springer, Berlin.
- ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* **11** R106. <https://doi.org/10.1186/gb-2010-11-10-r106>
- BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. [MR2091634 https://doi.org/10.1126/science.286.5439.509](https://doi.org/10.1126/science.286.5439.509)
- BARBER, R. F. and DRTON, M. (2015). High-dimensional Ising model selection with Bayesian information criteria. *Electron. J. Stat.* **9** 567–607. [MR3326135 https://doi.org/10.1214/15-EJS1012](https://doi.org/10.1214/15-EJS1012)
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. [MR1325392](https://doi.org/10.1111/j.1467-9868.1995.tb00331.x)
- BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. [MR0373208](https://doi.org/10.2307/2344138)
- BRAITHWAITE, A. T., MARRIOTT, H. M. and LAWRIE, A. (2018). Divergent roles for TRAIL in lung diseases. *Front. Med.* **5** 212. <https://doi.org/10.3389/fmed.2018.00212>
- CAI, T. T., LI, H., MA, J. and XIA, Y. (2019). Differential Markov random field analysis with an application to detecting differential microbial community networks. *Biometrika* **106** 401–416. [MR3949311 https://doi.org/10.1093/biomet/asz012](https://doi.org/10.1093/biomet/asz012)
- CHEN, J., BARDES, E. E., ARONOW, B. J. and JEGGA, A. G. (2009). ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37** W305–W311. <https://doi.org/10.1093/nar/gkp427>

Key words and phrases. Poisson graphical model, asthma genomics, RNA-seq data, asymptotic normality, bias correction, multiple testing.

- FORNO, E., WANG, T., QI, C., YAN, Q., XU, C.-J., BOUTAOUI, N., HAN, Y.-Y., WEEKS, D. E., JIANG, Y. et al. (2019). DNA methylation in nasal epithelium, atopy, and atopic asthma in children: A genome-wide study. *Lancet Respir. Med.* **7** 336–346.
- FORNO, E., ZHANG, R., JIANG, Y., KIM, S., YAN, Q., REN, Z., HAN, Y.-Y., BOUTAOUI, N., ROSSER, F. et al. (2020). Transcriptome-wide and differential expression network analyses of childhood asthma in nasal epithelium. *J. Allergy Clin. Immunol.* **146** 671–675.
- GIOVANNINI-CHAMI, L., MARCET, B., MOREILHON, C., CHEVALIER, B., ILLIE, M. I., LEBRIGAND, K., ROBBE-SERMESANT, K., BOURRIER, T., MICHIELS, J.-F. et al. (2012). Distinct epithelial gene expression phenotypes in childhood respiratory allergy. *Eur. Respir. J.* **39** 1197–1205.
- HADEBE, S., BROMBACHER, F. and BROWN, G. D. (2018). C-type lectins receptors in asthma. *Front. Immunol.* **9** 733.
- INOUE, D., RAVIKUMAR, P. and DHILLON, I. (2016). Square root graphical models: Multivariate generalizations of univariate exponential families that permit positive dependencies. In *International Conference on Machine Learning* 2445–2453.
- ISLAM, S., ZEISEL, A., JOOST, S., LA MANNO, G., ZAJAC, P., KASPER, M., LÖNNERBERG, P. and LINNARSSON, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11** 163–166.
- JANKOVÁ, J. and VAN DE GEER, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Stat.* **9** 1205–1229. [MR3354336 https://doi.org/10.1214/15-EJS1031](https://doi.org/10.1214/15-EJS1031)
- JANKOVÁ, J. and VAN DE GEER, S. (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *TEST* **26** 143–162. [MR3613609 https://doi.org/10.1007/s11749-016-0503-5](https://doi.org/10.1007/s11749-016-0503-5)
- JIA, B., XU, S., XIAO, G., LAMBA, V. and LIANG, F. (2017). Learning game regulatory networks from next generation sequencing data. *Biometrics* **73** 1221–1230. [MR3744536 https://doi.org/10.1111/biom.12682](https://doi.org/10.1111/biom.12682)
- KABESCH, M., SCHEDEL, M., CARR, D., WOITSCH, B., FRITZSCH, C., WEILAND, S. K. and VON MUTIUS, E. (2006). IL-4/IL-13 pathway genetics strongly influence serum IgE levels and childhood asthma. *J. Allergy Clin. Immunol.* **117** 269–274.
- LI, S., REN, Z., ZHANG, C.-H. and ZHOU, H. H. (2016). Asymptotic normality in estimation of large Ising graphical models. Unpublished Manuscript.
- LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* **41** 2948–2978. [MR3161453 https://doi.org/10.1214/13-AOS1169](https://doi.org/10.1214/13-AOS1169)
- LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. [MR2563983 https://doi.org/10.1214/09-AOS691](https://doi.org/10.1214/09-AOS691)
- LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. [MR3059084 https://doi.org/10.1214/12-AOS1037](https://doi.org/10.1214/12-AOS1037)
- NEWMAN, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* (3) **74** 036104. [MR2282139 https://doi.org/10.1103/PhysRevE.74.036104](https://doi.org/10.1103/PhysRevE.74.036104)
- PANDEY, G., PANDEY, O. P., ROGERS, A. J., AHSEN, M. E., HOFFMAN, G. E., RABY, B. A., WEISS, S. T., SCHADT, E. E. and BUNYAVANICH, S. (2018). A nasal brush-based classifier of asthma identified by machine learning analysis of nasal RNA sequence data. *Sci. Rep.* **8** 8826. <https://doi.org/10.1038/s41598-018-27189-4>
- POTAPINSKA, O. and DEMKOW, U. (2009). T lymphocyte apoptosis in asthma. *Eur. J. Med. Res.* **14** 192–195.
- RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1 -regularized logistic regression. *Ann. Statist.* **38** 1287–1319. [MR2662343 https://doi.org/10.1214/09-AOS691](https://doi.org/10.1214/09-AOS691)
- REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. [MR3346695 https://doi.org/10.1214/14-AOS1286](https://doi.org/10.1214/14-AOS1286)
- VALE, K. (2016). Targeting the JAK-STAT pathway in the treatment of Th2-high severe asthma. *Future Med. Chem.* **8** 405–419.
- VAN CROMBRUGGEN, K., VOGL, T., PÉREZ-NOVO, C., HOLTAPPELS, G. and BACHERT, C. (2016). Differential release and deposition of S100A8/A9 proteins in inflamed upper airway tissue. *Eur. Respir. J.* **47** 264–274.
- VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. [MR3224285 https://doi.org/10.1214/14-AOS1221](https://doi.org/10.1214/14-AOS1221)
- WANG, J. and KOLAR, M. (2016). Inference for high-dimensional exponential family graphical models. In *Artificial Intelligence and Statistics* 1042–1050.
- WANG, H., FITZPATRICK, M., WILSON, N. J., ANTHONY, D., READING, P. C., SATZKE, C., DUNNE, E. M., LICCIARDI, P. V., SEOW, H. J. et al. (2019). CSF3R/CD114 mediates infection-dependent transition to severe asthma. *J. Allergy Clin. Immunol.* **143** 785–788.

- WILLIAMS, J. W., FERREIRA, C. M., BLAINE, K. M., RAYON, C., VELÁZQUEZ, F., TONG, J., PETER, M. E. and SPERLING, A. I. (2018). Non-apoptotic Fas (CD95) signaling on T cells regulates the resolution of Th2-mediated inflammation. *Front. Immunol.* **9** 2521.
- YANG, X., ZHANG, B., MOLONY, C., CHUDIN, E., HAO, K., ZHU, J., GAEDIGK, A., SUVER, C., ZHONG, H. et al. (2010). Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. *Genome Res.* **20** 1020–1036.
- YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2013). On Poisson graphical models. In *Advances in Neural Information Processing Systems* 1718–1726.
- YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16** 3813–3847. [MR3450553](#)
- YANG, I. V., PEDERSEN, B. S., LIU, A. H., O’CONNOR, G. T., PILLAI, D., KATTAN, M., MISIAK, R. T., GRUCHALLA, R., SZEFLER, S. J. et al. (2017). The nasal methylome and childhood atopic asthma. *J. Allergy Clin. Immunol.* **139** 1478–1488.
- YU, M., KOLAR, M. and GUPTA, V. (2016). Statistical inference for pairwise graphical models using score matching. In *Advances in Neural Information Processing Systems* 2829–2837.
- ZHANG, Y., OUYANG, Z. and ZHAO, H. (2017). A statistical framework for data integration through graphical models with application to cancer genomics. *Ann. Appl. Stat.* **11** 161–184. [MR3634319](#) <https://doi.org/10.1214/16-AOAS998>
- ZHANG, R., REN, Z. and CHEN, W. (2018). SILGGM: An extensive R package for efficient statistical inference in large-scale gene networks. *PLoS Comput. Biol.* **14** e1006369.
- ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. [MR3153940](#) <https://doi.org/10.1111/rssb.12026>
- ZHANG, N.-Z., CHEN, X.-J., MU, Y.-H. and WANG, H. (2018). Identification of differentially expressed genes in childhood asthma. *Medicine* **97** e10861.
- ZHANG, R., REN, Z., CELEDÓN, J. C and CHEN, W. (2021). Supplement to “Inference of large modified poisson-type graphical models: Application to RNA-seq data in childhood atopic asthma studies.” <https://doi.org/10.1214/20-AOAS1413SUPPA>, <https://doi.org/10.1214/20-AOAS1413SUPPB>
- ZHAO, T., LIU, H., ROEDER, K., LAFFERTY, J. and WASSERMAN, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13** 1059–1062. [MR2930633](#)
- ZWIENER, I., FRISCH, B. and BINDER, H. (2014). Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS ONE* **9** e85150. <https://doi.org/10.1371/journal.pone.0085150>

TWO-WAY SPARSITY FOR TIME-VARYING NETWORKS WITH APPLICATIONS IN GENOMICS

BY THOMAS E. BARTLETT^{1,*}, IOANNIS KOSMIDIS^{2,3} AND RICARDO SILVA^{1,3}

¹Department of Statistics, University College London *thomas.bartlett.10@ucl.ac.uk

²Department of Statistics, University of Warwick, Coventry

³The Alan Turing Institute, London

We propose a novel way of modelling time-varying networks by inducing two-way sparsity on local models of node connectivity. This two-way sparsity separately promotes sparsity across time and sparsity across variables (within time). Separation of these two types of sparsity is achieved through a novel prior structure which draws on ideas from the Bayesian lasso and from copula modelling. We provide an efficient implementation of the proposed model via a Gibbs sampler, and we apply the model to data from neural development. In doing so, we demonstrate that the proposed model is able to identify changes in genomic network structure that match current biological knowledge. Such changes in genomic network structure can then be used by neurobiologists to identify potential targets for further experimental investigation.

REFERENCES

- AIBAR, S., GONZÁLEZ-BLAS, C. B., MOERMAN, T., IMRICHOVA, H., HULSELMANS, G., RAMBOW, F., MARINE, J.-C., GEURTS, P., AERTS, J. et al. (2017). SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **14** 1083.
- ALCAMO, E. A., CHIRIVELLA, L., DAUTZENBERG, M., DOBREVA, G., FARIÑAS, I., GROSSCHEDL, R. and MCCONNELL, S. K. (2008). Satb2 regulates callosal projection neuron identity in the developing cerebral cortex. *Neuron* **57** 364–377.
- ALEXANDER, R. P., KIM, P. M., EMONET, T. and GERSTEIN, M. B. (2009). Understanding modularity in molecular networks requires dynamics. *Sci. Signal.* **2** pe44.
- ALSANIE, W. F., PENNA, V., SCHACHNER, M., THOMPSON, L. H. and PARISH, C. L. (2017). Homophilic binding of the neural cell adhesion molecule CHL1 regulates development of ventral midbrain dopaminergic pathways. *Sci. Rep.* **7** 9368. <https://doi.org/10.1038/s41598-017-09599-y>
- ANDREWS, D. F. and MALLOWS, C. L. (1974). Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B* **36** 99–102. [MR0359122](https://doi.org/10.2307/2344682)
- BARTLETT, T. E., KOSMIDIS, I. and SILVA, R. (2021). Supplement to “Two-way sparsity for time-varying networks, with applications in genomics.” <https://doi.org/10.1214/20-AOAS1416SUPP>
- CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. [MR2650751 https://doi.org/10.1093/biomet/asq017](https://doi.org/10.1093/biomet/asq017)
- CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. [MR3375874 https://doi.org/10.1214/15-AOS1334](https://doi.org/10.1214/15-AOS1334)
- CECH, T. R. and STEITZ, J. A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157** 77–94. <https://doi.org/10.1016/j.cell.2014.03.008>
- CRANE, H. (2016). Dynamic random networks and their graph limits. *Ann. Appl. Probab.* **26** 691–721. [MR3476622 https://doi.org/10.1214/15-AAP1098](https://doi.org/10.1214/15-AAP1098)
- DURANTE, D. and DUNSON, D. B. (2016). Locally adaptive dynamic networks. *Ann. Appl. Stat.* **10** 2203–2232. [MR3592054 https://doi.org/10.1214/16-AOAS971](https://doi.org/10.1214/16-AOAS971)
- FAN, J., FENG, Y. and WU, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.* **3** 521–541. [MR2750671 https://doi.org/10.1214/08-AOAS215](https://doi.org/10.1214/08-AOAS215)
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

- 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 Dept. Minneapolis, MN.
- HEIDELBERGER, P. and WELCH, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Commun. ACM* **24** 233–245. [MR0611745 https://doi.org/10.1145/358598.358630](https://doi.org/10.1145/358598.358630)
- JOHN, A., BRYLKA, H., WIEGREFFE, C., SIMON, R., LIU, P., JÜTTNER, R., CRENSHAW, E. B., LUYTEN, F. P., JENKINS, N. A. et al. (2012). Bcl11a is required for neuronal morphogenesis and sensory circuit formation in dorsal spinal cord development. *Development* **139** 1831–1841.
- KALAITZIS, A., LAFFERTY, J., LAWRENCE, N. and ZHOU, S. (2013). The bigraphical lasso. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* 1229–1237.
- KHARCHENKO, P. V., SILBERSTEIN, L. and SCADDEN, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11** 740–742.
- KOLAR, M., SONG, L., AHMED, A. and XING, E. P. (2010). Estimating time-varying networks. *Ann. Appl. Stat.* **4** 94–123. [MR2758086 https://doi.org/10.1214/09-AOAS308](https://doi.org/10.1214/09-AOAS308)
- KYUNG, M., GILL, J., GHOSH, M. and CASELLA, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal.* **5** 369–411. [MR2719657 https://doi.org/10.1214/10-BA607](https://doi.org/10.1214/10-BA607)
- LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. The Clarendon Press, New York. [MR1419991](https://doi.org/10.1093/oxfordjbs.17.1.1)
- LÈBRE, S., BECQ, J., DEVAUX, F., STUMPF, M. P. H. and LELANDAIS, G. (2010). Statistical inference of the time-varying structure of gene-regulation networks. *BMC Syst. Biol.* **4** 130. <https://doi.org/10.1186/1752-0509-4-130>
- LI, H., RADFORD, J. C., RAGUSA, M. J., SHEA, K. L., MCKERCHER, S. R., ZAREMBA, J. D., SOUSSOU, W., NIE, Z., KANG, Y.-J. et al. (2008). Transcription factor MEF2C influences neural stem/progenitor cell differentiation and maturation in vivo. *Proc. Natl. Acad. Sci. USA* **105** 9397–9402.
- LINQING, Z., GUOHUA, J., HAOMING, L., XUELEI, T., JIANBING, Q. and MEILING, T. (2015). RUNX1T1 regulates the neuronal differentiation of radial glial cells from the rat hippocampus. *Stem Cells Translational Medicine* **4** 110–116.
- MATIAS, C. and MIELE, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1119–1141. [MR3689311 https://doi.org/10.1111/rssb.12200](https://doi.org/10.1111/rssb.12200)
- MATSUMOTO, H., KIRYU, H., FURUSAWA, C., KO, M. S. H., KO, S. B. H., GOUDA, N., HAYASHI, T. and NIKAIIDO, I. (2017). SCODE: An efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33** 2314–2321. <https://doi.org/10.1093/bioinformatics/btx194>
- MAYER, S., CHEN, J., VELMESHEV, D., MAYER, A., EZE, U. C., BHADURI, A., CUNHA, C. E., JUNG, D., ARJUN, A. et al. (2019). Multimodal single-cell analysis reveals physiological maturation in the developing human neocortex. *Neuron*.
- NOVERSHTERN, N., REGEV, A. and FRIEDMAN, N. (2011). Physical module networks: An integrative approach for reconstructing transcription regulation. *Bioinformatics* **27** i177–i185.
- NOWAKOWSKI, T. J., BHADURI, A., POLLEN, A. A., ALVARADO, B., MOSTAJO-RADJI, M. A., DI LULLO, E., HAEUSSLER, M., SANDOVAL-ESPINOSA, C., LIU, S. J. et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358** 1318–1323.
- PALLA, K., CARON, F. and TEH, Y. W. (2016). Bayesian nonparametrics for sparse dynamic networks. Preprint. Available at [arXiv:1607.01624](https://arxiv.org/abs/1607.01624).
- PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. [MR2524001 https://doi.org/10.1198/016214508000000337](https://doi.org/10.1198/016214508000000337)
- PATASKAR, A., JUNG, J., SMIALOWSKI, P., NOACK, F., CALEGARI, F., STRAUB, T. and TIWARI, V. K. (2016). NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program. *The EMBO Journal* **35** 24–45.
- PENSKY, M. (2019). Dynamic network models and graphon estimation. *Ann. Statist.* **47** 2378–2403. [MR3953455 https://doi.org/10.1214/18-AOS1751](https://doi.org/10.1214/18-AOS1751)
- PIERSON, E. and YAU, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16** 1–10.
- PIPER, M., BARRY, G., HAWKINS, J., MASON, S., LINDWALL, C., LITTLE, E., SARKAR, A., SMITH, A. G., MOLDRICH, R. X. et al. (2010). NFIA controls telencephalic progenitor cell differentiation through repression of the Notch effector Hes1. *J. Neurosci.* **30** 9127–9139.
- PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6** 7–11.
- QIU, P., SIMONDS, E. F., BENDALL, S. C., GIBBS JR, K. D., BRUGGNER, R. V., LINDERMAN, M. D., SACHS, K., NOLAN, G. P. and PLEVITIS, S. K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29** 886–891.

- REHFELD, F., MATICZKA, D., GROSSER, S., KNAUFF, P., ERAVCI, M., VIDA, I., BACKOFEN, R. and WULCZYN, F. G. (2018). The RNA-binding protein ARPP21 controls dendritic branching by functionally opposing the miRNA it hosts. *Nat. Commun.* **9** 1235. <https://doi.org/10.1038/s41467-018-03681-3>
- ROSENGREN, S. and TRAPMAN, P. (2019). A dynamic Erdős–Rényi graph model. *Markov Process. Related Fields* **25** 275–300. [MR3967544](https://doi.org/10.1017/S1446788719000054)
- SARKAR, P., CHAKRABARTI, D. and JORDAN, M. (2014). Nonparametric link prediction in large scale dynamic networks. *Electron. J. Stat.* **8** 2022–2065. [MR3273618](https://doi.org/10.1214/14-EJS943)
- SCHAEFER, A., MARGULIES, D. S., LOHMANN, G., GORGOLEWSKI, K. J., SMALLWOOD, J., KIEBEL, S. J. and VILLRINGER, A. (2014). Dynamic network participation of functional connectivity hubs assessed by resting-state fMRI. *Front. Human Neurosci.* **8** 195. <https://doi.org/10.3389/fnhum.2014.00195>
- SEKARA, V., STOPCZYNSKI, A. and LEHMANN, S. (2016). Fundamental structures of dynamic social networks. *Proc. Natl. Acad. Sci. USA* **113** 9977–9982.
- SHIMAMURA, K., UEKI, M., KAWANO, S. and KONISHI, S. (2019). Bayesian generalized fused lasso modeling via NEG distribution. *Comm. Statist. Theory Methods* **48** 4132–4153. [MR3976726](https://doi.org/10.1080/03610926.2018.1489056)
- SUVÀ, M. L., RHEINBAY, E., GILLESPIE, S. M., PATEL, A. P., WAKIMOTO, H., RABKIN, S. D., RIGGI, N., CHI, A. S., CAHILL, D. P. et al. (2014). Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell* **157** 580–594.
- TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. [MR2850205](https://doi.org/10.1214/11-AOS878)
- TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. [MR2136641](https://doi.org/10.1111/j.1467-9868.2005.00490.x)
- TORIYAMA, M., SHIMADA, T., KIM, K. B., MITSUBA, M., NOMURA, E., KATSUTA, K., SAKUMURA, Y., ROEPSTORFF, P. and INAGAKI, N. (2006). Shootin1: A protein involved in the organization of an asymmetric signal for neuronal polarization. *J. Cell Biol.* **175** 147–157.
- TRAPNELL, C., CACCHIARELLI, D., GRIMSBY, J., POKHAREL, P., LI, S., MORSE, M., LENNON, N. J., LIVAK, K. J., MIKKELSEN, T. S. et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32** 381–386.
- VAN DIJK, D., NAINYS, J., SHARMA, R., KATHAIL, P., CARR, A. J., MOON, K. R., MAZUTIS, L., WOLF, G., KRISHNASWAMY, S. et al. (2017). MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *BioRxiv* 111591.
- VAN DER PAS, S. L., SALOMOND, J.-B. and SCHMIDT-HIEBER, J. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electron. J. Stat.* **10** 976–1000. [MR3486423](https://doi.org/10.1214/16-EJS1130)
- XU, K. S. and HERO III, A. O. (2013). Dynamic stochastic blockmodels: Statistical models for time-evolving networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* 201–210. Springer.
- ZHANG, S., ZHAO, J. and ZHANG, X.-S. (2012). Common community structure in time-varying networks. *Phys. Rev. E* (3) **85** 056110.

A COMPOSITIONAL MODEL TO ASSESS EXPRESSION CHANGES FROM SINGLE-CELL RNA-SEQ DATA

BY XIUYU MA^{1,*}, KEEGAN KORTHAUER², CHRISTINA KENDZIORSKI³ AND
 MICHAEL A. NEWTON^{1,†}

¹Department of Statistics, University of Wisconsin–Madison, *watsonforfun@gmail.com; †newton@stat.wisc.edu

²Department of Statistics, University of British Columbia, keegan@stat.ubc.ca

³Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, kendzior@biostat.wisc.edu

On the problem of scoring genes for evidence of changes in the distribution of single-cell expression, we introduce an empirical Bayesian mixture approach and evaluate its operating characteristics in a range of numerical experiments. The proposed approach leverages cell-subtype structure revealed in cluster analysis in order to boost gene-level information on expression changes. Cell clustering informs gene-level analysis through a specially-constructed prior distribution over pairs of multinomial probability vectors; this prior meshes with available model-based tools that score patterns of differential expression over multiple subtypes. We derive an explicit formula for the posterior probability that a gene has the same distribution in two cellular conditions, allowing for a gene-specific mixture over subtypes in each condition. Advantage is gained by the compositional structure of the model not only in which a host of gene-specific mixture components are allowed but also in which the mixing proportions are constrained at the whole cell level. This structure leads to a novel form of information sharing through which the cell-clustering results support gene-level scoring of differential distribution. The result, according to our numerical experiments, is improved sensitivity compared to several standard approaches for detecting distributional expression changes.

REFERENCES

- ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* **11** R106. <https://doi.org/10.1186/gb-2010-11-10-r106>
- BACHER, R. and KENDZIORSKI, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **17** 63. <https://doi.org/10.1186/s13059-016-0927-y>
- BUETTNER, F., NATARAJAN, K. N., CASALE, F. P., PROSERPIO, V., SCIALDONE, A., THEIS, F. J., TEICHMANN, S. A., MARIONI, J. C. and STEGLE, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33** 155.
- CHEN, W., LI, Y., EASTON, J., FINKELSTEIN, D., WU, G. and CHEN, X. (2018). UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol.* **19** 70. <https://doi.org/10.1186/s13059-018-1438-9>
- CHU, L.-F., LENG, N., ZHANG, J., HOU, Z., MAMOTT, D., VEREIDE, D. T., CHOI, J., KENDZIORSKI, C., STEWART, R. et al. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* **17** 173. <https://doi.org/10.1186/s13059-016-1033-x>
- DAHL, D. B. (2009). Modal clustering in a class of product partition models. *Bayesian Anal.* **4** 243–264. MR2507363 <https://doi.org/10.1214/09-BA409>
- DARMANIS, S., SLOAN, S. A., CROOTE, D., MIGNARDI, M., CHERNIKOVA, S., SAMGHABABI, P., ZHANG, Y., NEFF, N., KOWARSKY, M. et al. (2017). Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep.* **21** 1399–1410. <https://doi.org/10.1016/j.celrep.2017.10.030>
- DELMANS, M. and HEMBERG, M. (2016). Discrete distributional differential expression (D3E)—A tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinform.* **17** 110. <https://doi.org/10.1186/s12859-016-0944-6>

Key words and phrases. Local false discovery rate, mixture model, empirical Bayes, clustering, double Dirichlet mixture.

- DENG, Q., RAMSKÖLD, D., REINIUS, B. and SANDBERG, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343** 193–196. <https://doi.org/10.1126/science.1245316>
- DOMINGUEZ, D., TSAI, Y.-H., GOMEZ, N., JHA, D. K., DAVIS, I. and WANG, Z. (2016). A high-resolution transcriptome map of cell cycle reveals novel connections between periodic genes and cancer. *Cell Res.* **26** 946.
- EFRON, B. (2007). Size, power and false discovery rates. *Ann. Statist.* **35** 1351–1377. [MR2351089 https://doi.org/10.1214/009053606000001460](https://doi.org/10.1214/009053606000001460)
- ENGEL, I., SEUMOIS, G., CHAVEZ, L., SAMANIEGO-CASTRUITA, D., WHITE, B., CHAWLA, A., MOCK, D., VIJAYANAND, P. and KRONENBERG, M. (2016). Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. *Nat. Immunol.* **17** 728.
- FINAK, G., MCDAVID, A., YAJIMA, M., DENG, J., GERSUK, V., SHALEK, A. K., SLICHTER, C. K., MILLER, H. W., MCEL RATH, M. J. et al. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16** 278. <https://doi.org/10.1186/s13059-015-0844-5>
- GUO, F., YAN, L., GUO, H., LI, L., HU, B., ZHAO, Y., YONG, J., HU, Y., WANG, X. et al. (2015). The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell* **161** 1437–1452. <https://doi.org/10.1016/j.cell.2015.05.015>
- HAGHVERDI, L., BUETTNER, F. and THEIS, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31** 2989–2998. <https://doi.org/10.1093/bioinformatics/btv325>
- HUANG, M., WANG, J., TORRE, E., DUECK, H., SHAFFER, S., BONASIO, R., MURRAY, J. I., RAJ, A., LI, M. et al. (2018). SAVER: Gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15** 539–542. <https://doi.org/10.1038/s41592-018-0033-z>
- JENSEN, S. T., ERKAN, I., ARNARDOTTIR, E. S. and SMALL, D. S. (2009). Bayesian testing of many hypotheses $\times$ many genes: A study of sleep apnea. *Ann. Appl. Stat.* **3** 1080–1101. [MR2750387 https://doi.org/10.1214/09-AOAS241](https://doi.org/10.1214/09-AOAS241)
- KAUFMAN, L. and ROUSSEEUW, P. (1987). *Clustering by Means of Medoids*. North-Holland, Amsterdam.
- KENDZIORSKI, C. M., NEWTON, M. A., LAN, H. and GOULD, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med.* **22** 3899–3914. <https://doi.org/10.1002/sim.1548>
- KIM, T., CHEN, I. R., LIN, Y., WANG, A. Y.-Y., YANG, J. Y. H. and YANG, P. (2019). Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief. Bioinform.* **20** 2316–2326. <https://doi.org/10.1093/bib/bby076>
- KISELEV, V. Y., KIRSCHNER, K., SCHAUB, M. T., ANDREWS, T., YIU, A., CHANDRA, T., NATARAJAN, K. N., REIK, W., BARAHONA, M. et al. (2017). SC3: Consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14** 483.
- KORTHAUER, K. D., CHU, L.-F., NEWTON, M. A., LI, Y., THOMSON, J., STEWART, R. and KENDZIORSKI, C. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* **17** 222. <https://doi.org/10.1186/s13059-016-1077-y>
- LANE, K., VAN VALEN, D., DEFELICE, M. M., MACKLIN, D. N., KUDO, T., JAIMOVICH, A., CARR, A., MEYER, T., PE’ER, D. et al. (2017). Measuring signaling and RNA-seq in the same cell links gene expression to dynamic patterns of NF- κ B activation. *Cell Syst.* **4** 458–469.e5. <https://doi.org/10.1016/j.cels.2017.03.010>
- LENG, N., DAWSON, J. A., THOMSON, J. A., RUOTTI, V., RISSMAN, A. I., SMITS, B. M. G., HAAG, J. D., GOULD, M. N., STEWART, R. M. et al. (2013). EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29** 1035–1043. <https://doi.org/10.1093/bioinformatics/btt087>
- LENG, N., CHU, L.-F., BARRY, C., LI, Y., CHOI, J., LI, X., JIANG, P., STEWART, R. M., THOMSON, J. A. et al. (2015). Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods* **12** 947.
- LI, F. and ALTIERI, D. C. (1999). The cancer antiapoptosis mouse *Survivin* gene. *Cancer Res.* **59** 3143.
- LIN, P., TROUP, M. and HO, J. W. K. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **18** 59. <https://doi.org/10.1186/s13059-017-1188-0>
- LITTLE, A. F., MAGGIONI, M. and MURPHY, J. M. (2017). Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms.
- LOVE, M. I., HUBER, W. and ANDERS, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15** 550. <https://doi.org/10.1186/s13059-014-0550-8>
- MA, X., KENDZIORSKI, C. and NEWTON, M. A. (2020). EBSeq: Improving mixing computations for multi-group differential expression analysis. *BioRxiv*. <https://doi.org/10.1101/2020.06.19.162180>

- MA, X., KORTHAUER, K., KENDZIORSKI, C. and NEWTON, M. A. (2021). Supplement to “A compositional model to assess expression changes from single-cell RNA-seq data.” <https://doi.org/10.1214/20-AOAS1423SUPPA>, <https://doi.org/10.1214/20-AOAS1423SUPPB>, <https://doi.org/10.1214/20-AOAS1423SUPPC>
- MARIONI, J. C. and ARENDT, D. (2017). How single-cell genomics is changing evolutionary and developmental biology. *Annu. Rev. Cell Dev. Biol.* **33** 537–553. <https://doi.org/10.1146/annurev-cellbio-100616-060818>
- MCDAVID, A., DENNIS, L., DANAHER, P., FINAK, G., KROUSE, M., WANG, A., WEBSTER, P., BEECHEM, J. and GOTTARDO, R. (2014). Modeling bi-modality improves characterization of cell cycle on gene expression in single cells. *PLoS Comput. Biol.* **10** e1003696.
- MURALIDHARAN, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *Ann. Appl. Stat.* **4** 422–438. [MR2758178 https://doi.org/10.1214/09-AOAS276](https://doi.org/10.1214/09-AOAS276)
- NAVIN, N. E. (2015). The first five years of single-cell cancer genomics and beyond. *Genome Res.* **25** 1499–1507. <https://doi.org/10.1101/gr.191098.115>
- NAWY, T. (2013). Single-cell sequencing. *Nat. Methods* **11** 18.
- 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. <https://doi.org/10.1093/biostatistics/5.2.155>
- PAPALEXI, E. and SATIJA, R. (2017). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev., Immunol.* **18** 35.
- PECCOUD, J. and YCART, B. (1995). Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48** 222–234. <https://doi.org/10.1006/tpbi.1995.1027>
- PIERSON, E. and YAU, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16** 241. <https://doi.org/10.1186/s13059-015-0805-z>
- RAJ, A. and VAN OUDENAARDEN, A. (2008). Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell* **135** 216–226. <https://doi.org/10.1016/j.cell.2008.09.050>
- RAY, S. and TURI, R. H. (2000). Determination of number of clusters in K-Means clustering and application in colour image segmentation.
- SHALEK, A. K., SATIJA, R., SHUGA, J., TROMBETTA, J. J., GENNERT, D., LU, D., CHEN, P., GERTNER, R. S., GAUBLOMME, J. T. et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510** 363.
- SOHR, S. and ENGELAND, K. (2008). RHAMM is differentially expressed in the cell cycle and downregulated by the tumor suppressor p53. *Cell Cycle* **7** 3448–3460. <https://doi.org/10.4161/cc.7.21.7014>
- SONESON, C. and ROBINSON, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15** 255–261. <https://doi.org/10.1038/nmeth.4612>
- STREHL, A. and GHOSH, J. (2003). Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3** 583–617. [MR1991087 https://doi.org/10.1162/153244303321897735](https://doi.org/10.1162/153244303321897735)
- TASIC, B., MENON, V., NGUYEN, T. N., KIM, T. K., JARSKY, T., YAO, Z., LEVI, B., GRAY, L. T., SORENSEN, S. A. et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19** 335.
- TRAPNELL, C., CACCHIARELLI, D., GRIMSBY, J., POKHAREL, P., LI, S., MORSE, M., LENNON, N. J., LIVAK, K. J., MIKKELSEN, T. S. et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32** 381–386. <https://doi.org/10.1038/nbt.2859>
- WAGNER, U. and TAUDS, A. (1986). A multivariate Polya model of brand choice and purchase incidence. *Mark. Sci.* **5** 219–244. <https://doi.org/10.1287/mksc.5.3.219>
- YAKOWITZ, S. J. and SPRAGINS, J. D. (1968). On the identifiability of finite mixtures. *Ann. Math. Stat.* **39** 209–214. [MR0224204 https://doi.org/10.1214/aoms/1177698520](https://doi.org/10.1214/aoms/1177698520)
- YIN, G. and MA, Y. (2013). Pearson-type goodness-of-fit test with bootstrap maximum likelihood estimation. *Electron. J. Stat.* **7** 412–427. [MR3020427 https://doi.org/10.1214/13-EJS773](https://doi.org/10.1214/13-EJS773)
- ZAPPIA, L., PHIPSON, B. and OSHLACK, A. (2017). Splatter: Simulation of single-cell RNA sequencing data. *Genome Biol.* **18** 174. <https://doi.org/10.1186/s13059-017-1305-0>

LARGE-SCALE MULTIPLE INFERENCE OF COLLECTIVE DEPENDENCE WITH APPLICATIONS TO PROTEIN FUNCTION

BY ROBERT JERNIGAN^{1,*}, KEJUE JIA^{1,†}, ZHAO REN² AND WEN ZHOU³

¹Department of Biochemistry, Biophysics, and Molecular Biology, Program of Bioinformatics and Computational Biology,
Iowa State University, *jernigan@iastate.edu; †kjia@iastate.edu

²Department of Statistics, University of Pittsburgh, zren@pitt.edu

³Department of Statistics, Colorado State University, riczw@stat.colostate.edu

Measuring the dependence of $k \geq 3$ random variables and drawing inference from such higher-order dependences are scientifically important yet challenging. Motivated here by protein coevolution with multivariate categorical features, we consider an information theoretic measure of higher-order dependence. The proposed collective dependence is a symmetrization of differential interaction information which generalizes the mutual information of a pair of random variables. We show that the collective dependence can be easily estimated and facilitates a test on the dependence of $k \geq 3$ random variables. Upon carefully exploring the null space of collective dependence, we devise a Classification-Assisted Large scaLe inference procedure to **DE**tect significant **k-COL**lective **DE**pendence among $d \geq k$ random variables, with the false discovery rate controlled. Finite sample performance of our method is examined via simulations. We apply this method to the multiple protein sequence alignment data to study the residue or position coevolution for two protein families, the elongation factor P family and the zinc knuckle family. We identify novel functional triplets of amino acid residues, whose contributions to the protein function are further investigated. These confirm that the collective dependence does yield additional information important for understanding the protein coevolution compared to the pairwise measures.

REFERENCES

- AFONNIKOV, D. A. and KOLCHANOV, N. A. (2004). CRASP: A program for analysis of coordinated substitutions in multiple alignments of protein sequences. *Nucleic Acids Res.* **32** W64–W68.
- BASHARIN, G. P. (1959). On a statistical estimate for the entropy of a sequence of independent random variables. *Theory Probab. Appl.* **4** 333–336. [MR0127457 https://doi.org/10.1137/1104033](https://doi.org/10.1137/1104033)
- BELL, A. J. (2003). The co-information lattices. In *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: IC 2003*.
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. [MR1325392](https://doi.org/10.1111/j.2517-6161.1995.tb00331.x)
- BERMAN, H., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T., WEISSIG, H., SHINDYALOV, I. and BOURNE, P. (2000). The protein data bank. *Nucleic Acids Res.* **28** 235–242.
- BERRETT, T. B., SAMWORTH, R. J. and YUAN, M. (2019). Efficient multivariate entropy estimation via k -nearest neighbour distances. *Ann. Statist.* **47** 288–318. [MR3909934 https://doi.org/10.1214/18-AOS1688](https://doi.org/10.1214/18-AOS1688)
- BUENO, R. and MAR, J. C. (2017). Changes in gene expression variability reveal a stable synthetic lethal interaction network in BRCA2-ovarian cancers. *Methods* **131** 74–82. <https://doi.org/10.1016/j.ymeth.2017.07.021>
- BURGER, L. and VAN NIMWEGEN, E. (2010). Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.* **6** e1000633, 18. [MR2601389 https://doi.org/10.1371/journal.pcbi.1000633](https://doi.org/10.1371/journal.pcbi.1000633)
- BUSLJE, C. M., TEPPA, E., DI DOMÉNICO, T., DELFINO, J. M. and NIELSEN, M. (2010). Networks of high mutual information define the structural proximity of catalytic sites: Implications for catalytic residue identification. *PLoS Comput. Biol.* **6** e1000978.
- CAI, T. T. and LIU, W. (2016). Large-scale multiple testing of correlations. *J. Amer. Statist. Assoc.* **111** 229–240. [MR3494655 https://doi.org/10.1080/01621459.2014.999157](https://doi.org/10.1080/01621459.2014.999157)

Key words and phrases. Collective dependence, false discovery rate, information theoretic measure, multiple testing, protein coevolution, structural biology.

- CAI, T. T., LI, H., LIU, W. and XIE, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100** 139–156. [MR3034329](https://doi.org/10.1093/biomet/ass058) <https://doi.org/10.1093/biomet/ass058>
- CHAKRABARTI, P. and BHATTACHARYYA, R. (2007). Geometry of nonbonded interactions involving planar groups in proteins. *Prog. Biophys. Mol. Biol.* **95** 83–137. <https://doi.org/10.1016/j.pbiomolbio.2007.03.016>
- CHAO, J. A., PATSKOVSKY, Y., ALMO, S. C. and SINGER, R. H. (2008). Structural basis for the coevolution of a viral RNA-protein complex. *Nat. Struct. Mol. Biol.* **15** 103–105.
- CIRIELLO, G., CERAMI, E., SANDER, C. and SCHULTZ, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22** 398–406. <https://doi.org/10.1101/gr.125567.111>
- COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley Interscience, Hoboken, NJ. [MR2239987](https://doi.org/10.1002/9781119223998)
- CUSICK, M. E., KLITGORD, N., VIDAL, M. and HILL, D. E. (2005). Interactome: Gateway into system biology. *Hum. Mol. Genet.* **14** R171–R181.
- DARWIN, C. (1859). *The Origin of Species by Means of Natural Selection*. Modern Lib.
- DE GUZMAN, R. N., WU, Z. R., STALLING, C. C., PAPPALARDO, L., BORER, P. N. and SUMMERS, M. F. (1998). Structure of the HIV-1 nucleocapsid protein bound to the SL3 Ψ -RNA recognition element. *Science* **279** 384–388.
- DE JUAN, D., PAZOS, F. and VALENCIA, A. (2015). Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **14** 249–261.
- DUDOIT, S. and VAN DER LAAN, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics. Springer Series in Statistics*. Springer, New York. [MR2373771](https://doi.org/10.1007/978-0-387-49317-6) <https://doi.org/10.1007/978-0-387-49317-6>
- DUNN, S. D., WAHL, L. M. and GLOOR, G. B. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24** 333–340.
- FIGLIUZZI, M., BARRAT-CHARLAIX, P. and WEIGT, M. (2018). How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol. Biol. Evol.* **35** 1018–1027. <https://doi.org/10.1093/molbev/msy007>
- FINN, R. D., TATE, J., MISTRY, J., COGGILL, P. C., SAMMUT, S. J., HOTZ, H. R., CERIC, G., FORSLUND, K., EDDY, S. R. et al. (2008). The Pfam protein families database. *Nucleic Acids Res.* **36** 281–288.
- FLEURET, F. (2004). Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* **5** 1531–1555. [MR2248026](https://doi.org/10.1162/jmlr.2004.5.1531)
- FRASER, H. B., HIRSH, A. E., WALL, D. P. and EISEN, M. B. (2004). Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci. USA* **101** 9033–9038.
- GALAS, D. J., NYKTER, M., CARTER, G. W., PRICE, N. D. and SHMULEVICH, I. (2010). Biological information as set-based complexity. *IEEE Trans. Inf. Theory* **56** 667–677. [MR2724399](https://doi.org/10.1109/TIT.2009.2037046) <https://doi.org/10.1109/TIT.2009.2037046>
- GALAS, D. J., SAKHANENKO, N. A., SKUPIN, A. and IGNAC, T. (2014). Describing the complexity of systems: Multivariable “set complexity” and the information basis of systems biology. *J. Comput. Biol.* **21** 118–140. [MR3164643](https://doi.org/10.1089/cmb.2013.0039) <https://doi.org/10.1089/cmb.2013.0039>
- HALABI, N., RIVOIRE, O., LEIBLER, S. and RANGANATHAN, R. (2009). Protein sectors: Evolutionary units of three-dimensional structure. *Cell* **138** 774–786.
- HANAWA-SUETSUGU, K., SEKINE, S., SAKAI, H., HORI-TAKEMOTO, C., TERADA, T., UNZAI, S., TAME, J. R., KURAMITSU, S., SHIROUZU, M. et al. (2004). Crystal structure of elongation factor P from *Thermus thermophilus* HB8. *Proc. Natl. Acad. Sci. USA* **101** 9595–9600.
- HOPF, T. A., INGRAHAM, J. B., POELWIJK, F. J., SCHÄRFE, C. P. I., SPRINGER, M., SANDER, C. and MARKS, D. S. (2017). Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35** 128–135. <https://doi.org/10.1038/nbt.3769>
- JERNIGAN, R., JIA, K., REN, Z. and ZHOU, W. (2021). Supplement to “Large-scale multiple inference of collective dependence with applications to protein function.” <https://doi.org/10.1214/20-AOAS1431SUPP>
- JIA, K. and JERNIGNA, R. L. (2018). SeqStruct: A new amino acid similarity matrix based on sequence correlations and structural contacts yields sequence-structure congruence. *BioRxiv*. <https://doi.org/10.1101/268904>
- JONES, D. J., BUCHAN, D. W., COZZETTO, D. and PONTIL, M. (2011). PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28** 184–190.
- KLAMT, S., HAUS, U.-U. and THEIS, F. (2009). Hypergraphs and cellular networks. *PLoS Comput. Biol.* **5** e1000385, 6. [MR2516078](https://doi.org/10.1371/journal.pcbi.1000385) <https://doi.org/10.1371/journal.pcbi.1000385>
- KORBER, B. T., FARBER, R. M., WOLPERT, D. H. and LAPEDES, A. S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proc. Natl. Acad. Sci. USA* **90** 7176–7180.
- KOZACHENKO, L. F. and LEONENKO, N. N. (1987). Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii* **23** 9–16.

- LI, A. and BARBER, R. F. (2019). Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 45–74. [MR3904779](#)
- LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* **41** 2948–2978. [MR3161453](#) <https://doi.org/10.1214/13-AOS1169>
- LOCKLESS, S. W. and RANGANATHAN, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286** 295–299.
- MARTIN, L., GLOOR, G. B., DUNN, S. and WAHL, L. M. (2005). Using information theory to search for co-evolving residues in proteins. *Bioinformatics* **21** 4116–4124.
- MCGILL, W. J. (1954). Multivariate information transmission. *Psychometrika* **19** 97–116.
- MILLER, G. (1955). Note on the bias of information estimates. *Inf. Theory Psychol. Probl. Methods* **II-B** 95–100.
- MISHRA, S. K. and JERNIGNA, R. L. (2018). Protein dynamic communities from elastic network models align closely to the communities defined by molecular dynamics. *PLoS ONE* **13** e0199225.
- MORCOS, F., PAGNANI, A., LUNT, B., BERTOLINO, A., MARKS, D. S., SANDER, C., ZECCHINA, R., ONUCHIC, J. N., HWA, T. et al. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA* **108** 1293–1301.
- NELSON, D. L. and COX, M. M. (2005). *Principles of Biochemistry*, 4th ed. W.H. Freeman, New York.
- OCHOA, D. and PAZOS, F. (2014). Practical aspects of protein coevolution. *Front. Cell Dev. Biol.* **2** 14. <https://doi.org/10.3389/fcell.2014.00014>
- ONKEN, A., DRAGOI, V. and OBERMAYER, K. (2012). A maximum entropy test for evaluating higher-order correlations in spike counts. *PLoS Comput. Biol.* **8** e1002539, 12. [MR2958407](#) <https://doi.org/10.1371/journal.pcbi.1002539>
- PANINSKI, L. (2003). Estimation of entropy and mutual information. *Neural Comput.* **15** 1191–1253.
- REYNOLDS, K. A., MCLAUGHLIN, R. N. and RANGANATHAN, R. (2011). Hot spots for allosteric regulation on protein surfaces. *Cell* **147** 1564–1575.
- SLATKIN, M. (2008). Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9** 477–485.
- STAUDE, B., GRÜN, S. and ROTTER, S. (2010). Higher-order correlations in non-stationary parallel spike trains: Statistical modeling and inference. *Front. Comput. Neurosci.* **4** 16. <https://doi.org/10.3389/fncom.2010.00016>
- STAUDE, B., ROTTER, S. and GRÜN, S. (2010). CuBIC: Cumulant based inference of higher-order correlations in massively parallel spike trains. *J. Comput. Neurosci.* **29** 327–350. [MR2721349](#) <https://doi.org/10.1007/s10827-009-0195-x>
- SUN, Y., ZHANG, N. R. and OWEN, A. B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *Ann. Appl. Stat.* **6** 1664–1688. [MR3058679](#) <https://doi.org/10.1214/12-AOAS561>
- THOMPSON, J. N. (1994). *The Coevolutionary Process*. Univ. Chicago Press, Chicago, IL.
- VEJMEĽKA, M. and PALUŠ, M. (2008). Inferring the directionality of coupling with conditional mutual information. *Phys. Rev. E* (3) **77** 026214, 12. [MR2453293](#) <https://doi.org/10.1103/PhysRevE.77.026214>
- VESTERGAARD, B., VAN, L. B., ANDERSEN, G. R., NYBORG, J., BUCKINGHAM, R. H. and KJELDGAARD, M. (2001). Bacterial polypeptide release factor RF2 is structurally distinct from eukaryotic eRF1. *Mol. Cell* **8** 1375–1382.
- WANG, Y. X. R., JIANG, K., FELDMAN, L. J., BICKEL, P. J. and HUANG, H. (2015). Inferring gene-gene interactions and functional modules using sparse canonical correlation analysis. *Ann. Appl. Stat.* **9** 300–323. [MR3341117](#) <https://doi.org/10.1214/14-AOAS792>
- WEIGT, M., WHITE, R. A., SZURMANT, H., HOCH, J. A. and HWA, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. USA* **106** 67–72.
- WELLS, J. A. and MCCLENDON, C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* **450** 1001–1009.
- XIA, Y., CAI, T. and CAI, T. T. (2018). Multiple testing of submatrices of a precision matrix with applications to identification of between pathway interactions. *J. Amer. Statist. Assoc.* **113** 328–339. [MR3803468](#) <https://doi.org/10.1080/01621459.2016.1251930>
- YEUNG, R. W. (1991). A new outlook on Shannon’s information measures. *IEEE Trans. Inf. Theory* **37** 466–474. [MR1145812](#) <https://doi.org/10.1109/18.79902>
- YIN, J. and LI, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5** 2630–2650. [MR2907129](#) <https://doi.org/10.1214/11-AOAS494>
- YUAN, M., LIU, R., FENG, Y. and SHANG, Z. (2018). Testing community structures for hypergraphs. Preprint. Available at [arXiv:1810.04617](https://arxiv.org/abs/1810.04617).
- ZHANG, R., REN, Z. and CHEN, W. (2018). SILGGM: An extensive R package for efficient statistical inference in large-scale gene networks. *PLoS Comput. Biol.* **14** e1006369.
- ZHI, W., MINTURN, J., RAPPAPORT, E., BRODEUR, G. and LI, H. (2013). Network-based analysis of multivariate gene expression data. *Methods Mol. Biol.* **972** 121–139.

A BAYESIAN NONPARAMETRIC MODEL FOR INFERRING SUBCLONAL POPULATIONS FROM STRUCTURED DNA SEQUENCING DATA

BY SHAI HE^{**1,*}, AARON SCHEIN^{**2}, VISHAL SARSANI^{1,†} AND PATRICK FLAHERTY^{1,‡}

¹*Department of Mathematics and Statistics, University of Massachusetts Amherst, *shaihe@math.umass.edu;
†vsarsani@umass.edu; ‡pflaherty@umass.edu*

²*Data Science Institute, Columbia University, aaron.schein@columbia.edu*

There are distinguishing features or “hallmarks” of cancer that are found across tumors, individuals and types of cancer, and these hallmarks can be driven by specific genetic mutations. Yet within a single tumor there is often extensive genetic heterogeneity as evidenced by single-cell and bulk DNA sequencing data. The goal of this work is to jointly infer the underlying genotypes of tumor subpopulations and the distribution of those subpopulations in individual tumors by integrating single-cell and bulk sequencing data. Understanding the genetic composition of the tumor at the time of treatment is important in the personalized design of targeted therapeutic combinations and monitoring for possible recurrence after treatment.

We propose a hierarchical Dirichlet process mixture model that incorporates the correlation structure induced by a structured sampling arrangement, and we show that this model improves the quality of inference. We develop a representation of the hierarchical Dirichlet process prior as a Gamma–Poisson hierarchy, and we use this representation to derive a fast Gibbs sampling inference algorithm using the augment-and-marginalize method. Experiments with simulation data show that our model outperforms standard numerical and statistical methods for decomposing admixed count data. Analyses of real acute lymphoblastic leukemia cancer sequencing dataset shows that our model improves upon state-of-the-art bioinformatic methods. An interpretation of the results of our model on this real dataset reveals comutated loci across samples.

REFERENCES

- ALDOUS, D. J. (1985). Exchangeability and related topics. In *École D’été de Probabilités de Saint-Flour, XIII—1983. Lecture Notes in Math.* **1117** 1–198. Springer, Berlin. [MR0883646 https://doi.org/10.1007/BFb0099421](https://doi.org/10.1007/BFb0099421)
- ALIZADEH, A. A., ARANDA, V., BARDELLI, A., BLANPAIN, C., BOCK, C., BOROWSKI, C., CALDAS, C., CALIFANO, A., DOHERTY, M. et al. (2015). Toward understanding and exploiting tumor heterogeneity. *Nat. Med.* <https://doi.org/10.1038/nm.3915>
- ALKAN, C., COE, B. P. and EICHLER, E. E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12** 363–376. <https://doi.org/10.1038/nrg2958>
- ANDOR, N., GRAHAM, T. A., JANSEN, M., XIA, L. C., AKTIPIS, C. A., PETRITSCH, C., JI, H. P. and MALEY, C. C. (2016). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22** 105–113. <https://doi.org/10.1038/nm.3984>
- ARAN, D., SIROTA, M. and BUTTE, A. J. (2015). Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**. <https://doi.org/10.1038/ncomms9971>
- BEERENWINKEL, N., SCHWARZ, R. F., GERSTUNG, M. and MARKOWETZ, F. (2015). Cancer evolution: Mathematical models and computational inference. *Syst. Biol.* **64** e1–e25. <https://doi.org/10.1093/sysbio/syu081>
- BONAVIA, R., INDA, M. D. M., CAVENEE, W. K. and FURNARI, F. B. (2011). Heterogeneity maintenance in glioblastoma: A social network. *Cancer Res.* <https://doi.org/10.1158/0008-5472.CAN-11-0153>
- BUDCZIES, J., PFARR, N., STENZINGER, A., TREUE, D., ENDRIIS, V., ISMAEEL, F., BANGEMANN, N., BLOHMER, J.-U., DIETEL, M. et al. (2016). Ioncopy: A novel method for calling copy number alterations in amplicon sequencing data including significance assessment. *Oncotarget* **7** 13236–13247. <https://doi.org/10.18632/oncotarget.7451>

Key words and phrases. Bayesian nonparametric, augment-and-marginalize, tumor heterogeneity, Dirichlet process mixture, DNA sequencing.

- CAMPBELL, P. J., PLEASANCE, E. D., STEPHENS, P. J., DICKS, E., RANCE, R., GOODHEAD, I., FOLLOWS, G. A., GREEN, A. R., FUTREAL, P. A. et al. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl. Acad. Sci. USA* **105** 13081–13086. <https://doi.org/10.1073/pnas.0801523105>
- CHOWELL, D., NAPIER, J., GUPTA, R., ANDERSON, K. S., MALEY, C. C. and WILSON SAYRES, M. A. (2018). Modeling the subclonal evolution of cancer cell populations. *Cancer Res.* **78** 830–839. <https://doi.org/10.1158/0008-5472.CAN-17-1229>
- DESHWAR, A. G., VEMBU, S., YUNG, C. K., JANG, G. H., STEIN, L. and MORRIS, Q. (2015). PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16** 35. <https://doi.org/10.1186/s13059-015-0602-8>
- ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. <https://doi.org/10.1080/01621459.1995.10476550>
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. <https://doi.org/10.1214/aos/1176342360>
- FISCHER, A., VÁZQUEZ-GARCÍA, I., ILLINGWORTH, C. J. R. and MUSTONEN, V. (2014). High-definition reconstruction of clonal composition in cancer. *Cell Rep.* **7** 1740–1752. <https://doi.org/10.1016/j.celrep.2014.04.055>
- GAWAD, C., KOH, W. and QUAKE, S. R. (2014). Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc. Natl. Acad. Sci. USA* **111** 17947–17952. <https://doi.org/10.1073/pnas.1420822111>
- GAWAD, C., KOH, W. and QUAKE, S. R. (2016). Single-cell genome sequencing: Current state of the science. *Nat. Rev. Genet.* 175–188. <https://doi.org/10.1038/nrg.2015.16>
- GERLINGER, M., ROWAN, A. J., HORSWELL, S., LARKIN, J., ENDESFELDER, D., GRONROOS, E., MARTINEZ, P., MATTHEWS, N., STEWART, A. et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366** 883–892. <https://doi.org/10.1056/nejmoa1113205>
- GOODWIN, S., MCPHERSON, J. D. and MCCOMBIE, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17** 333–351. <https://doi.org/10.1038/nrg.2016.49>
- GREEN, P. J. and RICHARDSON, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Stat.* **28** 355–375. <https://doi.org/10.1111/1467-9469.00242>
- HANAHAN, D. and WEINBERG, R. A. (2011). Hallmarks of cancer: The next generation. *Cell* **144** 646–674. <https://doi.org/10.1016/j.cell.2011.02.013>
- HE, S., SCHEIN, A., SARSANI, V. and FLAHERTY, P. (2021). Supplement to “A Bayesian nonparametric model for inferring subclonal populations from structured DNA sequencing data.” <https://doi.org/10.1214/20-AOAS1434SUPP>
- ISHWARAN, H. and ZAREPOUR, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87** 371–390. <https://doi.org/10.1093/biomet/87.2.371>
- ISHWARAN, H. and ZAREPOUR, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canad. J. Statist.* **30** 269–283. <https://doi.org/10.2307/3315951>
- KALISKY, T. and QUAKE, S. R. (2011). Single-cell genomics. *Nat. Methods* **8** 311–314. <https://doi.org/10.1038/nmeth0411-311>
- KYROCHRISTOS, I. D., ZIOGAS, D. E., GOUSSIA, A., GLANTZOUNIS, G. K. and ROUKOS, D. H. (2019). Bulk and single-cell next-generation sequencing: Individualizing treatment for colorectal cancer. *Cancers* **11**. <https://doi.org/10.3390/cancers11111809>
- LARSON, N. B. and FRIDLEY, B. L. (2013). PurBayes: Estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics* **29** 1888–1889. <https://doi.org/10.1093/bioinformatics/btt293>
- LEE, J., MÜLLER, P., GULUKOTA, K. and JI, Y. (2015). A Bayesian feature allocation model for tumor heterogeneity. *Ann. Appl. Stat.* **9** 621–639. <https://doi.org/10.1214/15-AOAS817>
- LI, H. and DURBIN, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26** 589–595. <https://doi.org/10.1093/bioinformatics/btp698>
- LOEB, L. A., KOHRN, B. F., LOUBET-SENEAR, K. J., DUNN, Y. J., AHN, E. H., O’SULLIVAN, J. N., SALK, J. J., BRONNER, M. P. and BECKMAN, R. A. (2019). Extensive subclonal mutational diversity in human colorectal cancer and its significance. *Proc. Natl. Acad. Sci. USA* **116** 26863–26872. <https://doi.org/10.1073/pnas.1910301116>
- LUKACS, E. (1955). A characterization of the gamma distribution. *Ann. Math. Stat.* **26** 319–324. <https://doi.org/10.1214/aoms/1177728549>
- MARASS, F., MOULIERE, F., YUAN, K., ROSENFELD, N. and MARKOWETZ, F. (2016). A phylogenetic latent feature model for clonal deconvolution. *Ann. Appl. Stat.* **10** 2377–2404. <https://doi.org/10.1214/16-AOAS986>

- MARTINCORENA, I. and CAMPBELL, P. J. (2015). Somatic mutation in cancer and normal cells. *Science* **349** 1483–1489. <https://doi.org/10.1126/science.aab4082>
- MARUSYK, A., ALMENDRO, V. and POLYAK, K. (2012). Intra-tumour heterogeneity: A looking glass for cancer? *Nat. Rev. Cancer*. <https://doi.org/10.1038/nrc3261>
- MILLER, J. W. and HARRISON, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems* 26 (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 199–206.
- MILLER, J. W. and HARRISON, M. T. (2014). Inconsistency of Pitman–Yor process mixtures for the number of components. *J. Mach. Learn. Res.* **15** 3333–3370. [MR3277163](https://doi.org/10.1162/jmlr.2014.15.1)
- MILLER, C. A., WHITE, B. S., DEES, N. D., GRIFFITH, M., WELCH, J. S., GRIFFITH, O. L., VIJ, R., TOMASSON, M. H., GRAUBERT, T. A. et al. (2014). SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.* **10** e1003665. <https://doi.org/10.1371/journal.pcbi.1003665>
- MÜLLER, P., QUINTANA, F. and ROSNER, G. (2004). A method for combining inference across related nonparametric Bayesian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 735–749. [MR2088779](https://doi.org/10.1111/j.1467-9868.2004.05564.x) <https://doi.org/10.1111/j.1467-9868.2004.05564.x>
- NAVIN, N. E. (2015). The first five years of single-cell cancer genomics and beyond. *Genome Res.* **25** 1499–1507. <https://doi.org/10.1101/gr.191098.115>
- NEAL, R. M. (1992). Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods*. https://doi.org/10.1007/978-94-017-2219-3_14
- NEUMANN, M., SEEHAWER, M., SCHLEE, C., VOSBERG, S., HEESCH, S., VON DER HEIDE, E. K., GRAF, A., KREBS, S., BLUM, H. et al. (2014). FAT1 expression and mutations in adult acute lymphoblastic leukemia. *Blood Cancer Journal* **4**.
- NOWELL, P. (1976). The clonal evolution of tumor cell populations. *Science* **194** 23–28. <https://doi.org/10.1126/science.959840>
- OH, J. H., JANG, S. J., KIM, J., SOHN, I., LEE, J. Y., CHO, E. J., CHUN, S. M. and SUNG, C. O. (2020). Spontaneous mutations in the single TTN gene represent high tumor mutation burden. *Npj Genomic Medicine*. <https://doi.org/10.1038/s41525-019-0107-6>
- PAISLEY, J. (2020). A tutorial on the Dirichlet process for engineers.
- PREDINA, J., ERUSLANOV, E., JUDY, B., KAPOOR, V., CHENG, G., WANG, L.-C., SUN, J., MOON, E. K., FRIDLINDER, Z. G. et al. (2013). Changes in the local tumor microenvironment in recurrent cancers may explain the failure of vaccines after surgery. *Proc. Natl. Acad. Sci. USA* **110** E415–424. <https://doi.org/10.1073/pnas.1211850110>
- RABADAN, R., BHANOT, G., MARSILIO, S., CHIORAZZI, N., PASQUALUCCI, L. and KHIABANIAN, H. (2018). On statistical modeling of sequencing noise in high depth data to assess tumor evolution. *J. Stat. Phys.* **172** 143–155. [MR3810540](https://doi.org/10.1007/s10955-017-1945-1) <https://doi.org/10.1007/s10955-017-1945-1>
- RASMUSSEN, C. E. (2000). The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems*.
- RIESTER, M., SINGH, A. P., BRANNON, A. R., YU, K., CAMPBELL, C. D., CHIANG, D. Y. and MORRISSEY, M. P. (2016). PureCN: Copy number calling and SNV classification using targeted short read sequencing. *Source Code for Biology and Medicine* **11**. <https://doi.org/10.1186/s13029-016-0060-z>
- ROTH, A., KHATTRA, J., YAP, D., WAN, A., LAKS, E., BIELE, J., HA, G., APARICIO, S., BOUCHARD-CÔTÉ, A. et al. (2014). Pyclone: Statistical inference of clonal population structure in cancer. *Nat. Methods* **11** 396–398. <https://doi.org/10.1038/nmeth.2883>
- RUSSNES, H. G., NAVIN, N., HICKS, J. and BORRESEN-DALE, A.-L. (2011). Insight into the heterogeneity of breast cancer through next-generation sequencing. *The Journal of Clinical Investigation* **121** 3810–3818. <https://doi.org/10.1172/JCI57088>
- SENGUPTA, S., WANG, J., LEE, J., MULLER, P., GULUKOTA, K., BANERJEE, A. and JI, Y. (2015). Bayclone: Bayesian nonparametric inference of tumor subclones using NGS data. In *Proceedings of the Pacific Symposium on Biocomputing* 467–478. [https://doi.org/9789814644730_0044\[prii\]](https://doi.org/9789814644730_0044[prii])
- SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. [MR1309433](https://doi.org/10.1111/j.1467-9868.2004.05564.x)
- STRATTON, M. R., CAMPBELL, P. J. and FUTREAL, P. A. (2009). The cancer genome. *Nature* **458** 719–724. <https://doi.org/10.1038/nature07943>
- SU, X., ZHANG, L., ZHANG, J., MERIC-BERNSTAM, F. and WEINSTEIN, J. N. (2012). Purityest: Estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics* **28** 2265–2266. <https://doi.org/10.1093/bioinformatics/bts365>
- 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](https://doi.org/10.1198/016214506000000302) <https://doi.org/10.1198/016214506000000302>

- TREUTLEIN, B., BROWNFIELD, D. G., WU, A. R., NEFF, N. F., MANTALAS, G. L., ESPINOZA, F. H., DE-SAI, T. J., KRASNOW, M. A. and QUAKE, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509** 371–375. <https://doi.org/10.1038/nature13173>
- VOGELSTEIN, B. and KINZLER, K. W. (2004). Cancer genes and the pathways they control. *Nat. Med.* **10** 789–799. <https://doi.org/10.1038/nm1087>
- ZAFAR, H., WANG, Y., NAKHLEH, L., NAVIN, N. and CHEN, K. (2016). Monovar: Single-nucleotide variant detection in single cells. *Nat. Methods* 505–507. <https://doi.org/10.1038/nmeth.3835>
- ZHANG, L., DONG, X., LEE, M., MASLOV, A. Y., WANG, T. and VIJG, J. (2019). Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc. Natl. Acad. Sci. USA* **116** 9014–9019. <https://doi.org/10.1073/pnas.1902510116>
- ZHOU, M. and CARIN, L. (2012). Augment-and-conquer negative binomial processes. In *Advances in Neural Information Processing Systems* 2546–2554.
- ZHOU, M. and CARIN, L. (2015). Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*
- ZHOU, M., HANNAH, L., DUNSON, D. and CARIN, L. (2012). Beta-negative binomial process and Poisson factor analysis. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* (N. D. Lawrence and M. Girolami, eds.). *Proceedings of Machine Learning Research* **22** 1462–1471. PMLR, La Palma, Canary Islands.
- ZHOU, T., SENGUPTA, S., MÜLLER, P. and JI, Y. (2019). TreeClone: Reconstruction of tumor subclone phylogeny based on mutation pairs using next generation sequencing data. *Ann. Appl. Stat.* **13** 874–899. MR3963556 <https://doi.org/10.1214/18-AOAS1224>

PREDICTION OF THE NASH THROUGH PENALIZED MIXTURE OF LOGISTIC REGRESSION MODELS

BY MARIE MORVAN^{1,*}, EMILIE DEVIJVER², MADISON GIACOFI^{1,†} AND VALÉRIE MONBET^{1,‡}

¹Université Rennes, CNRS, IRMAR—UMR 6625, *marie.morvan@univ-rennes1.fr; †joyce.giacofi@univ-rennes2.fr; ‡valerie.monbet@univ-rennes1.fr

²Université Grenoble Alpes, Inria, CNRS, Grenoble INP, emilie.devijver@univ-grenoble-alpes.fr

In this paper an appropriate and interpretable diagnosis statistical model is proposed to predict Nonalcoholic Steatohepatitis (NASH) from near infrared spectrometry data. In this disease, unknown patients' profiles are expected to lead to a different diagnosis. The model has then to take into account the heterogeneity of the data and the dimension of the spectrometric data.

To this end, we propose to fit a mixture model on the joint distribution of the diagnostic binary variable and the covariates selected in the spectra. The penalized maximum likelihood estimator is considered. In practice, a twofold penalty on both regression coefficients and covariance parameters is imposed. Automatic selection criteria, such as the AIC and BIC, are used to select the amount of shrinkage and the number of clusters. The performance of the overall procedure is evaluated by a simulation study, and its application on the NASH data set is analyzed. The model leads to better prediction performance than competitive methods and provides highly interpretable results.

REFERENCES

- AHONEN, I., NEVALAINEN, J. and LAROCQUE, D. (2019). Prediction with a flexible finite mixture-of-regressions. *Comput. Statist. Data Anal.* **132** 212–224. [MR3913145 https://doi.org/10.1016/j.csda.2018.01.012](https://doi.org/10.1016/j.csda.2018.01.012)
- ANTY, R., IANNELLI, A., PATOURAUX, S., BONNAFOUS, S., LAVALLARD, V., SENNI-BURATTI, M., BEN AMOR, I., STACCINI-MYX, A., SAINT-PAUL, M. C. et al. (2010). A new composite model including metabolic syndrome, alanine aminotransferase and cytokeratin-18 for the diagnosis of non-alcoholic steatohepatitis in morbidly obese patients. *Aliment. Pharmacol. Ther.* **32** 1315–1322.
- BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 719–725.
- BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput. Statist. Data Anal.* **41** 561–575. [MR1968069 https://doi.org/10.1016/S0167-9473\(02\)00163-9](https://doi.org/10.1016/S0167-9473(02)00163-9)
- BOUGEARD, S., ABDI, H., SAPORTA, G. and NIANG, N. (2018). Clusterwise analysis for multiblock component methods. *Adv. Data Anal. Classif.* **12** 285–313. [MR3829278 https://doi.org/10.1007/s11634-017-0296-8](https://doi.org/10.1007/s11634-017-0296-8)
- BRESLOW, N. E. (1996). Statistics in epidemiology: The case-control study. *J. Amer. Statist. Assoc.* **91** 14–28. [MR1394064 https://doi.org/10.2307/2291379](https://doi.org/10.2307/2291379)
- 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. With discussion. [MR0501537](https://doi.org/10.2307/2344607)
- FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. [MR1946581 https://doi.org/10.1198/016214501753382273](https://doi.org/10.1198/016214501753382273)
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.
- GREEN, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **52** 443–452. [MR1086796](https://doi.org/10.2307/2346178)
- GRÜN, B. and LEISCH, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Comput. Statist. Data Anal.* **51** 5247–5252. [MR2370869 https://doi.org/10.1016/j.csda.2006.08.014](https://doi.org/10.1016/j.csda.2006.08.014)

Key words and phrases. Mixture regression model, prediction, variable selection, heterogeneous data, spectrometry data.

- HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations. Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. [MR3616141](#)
- HOSHIKAWA, T. (2013). Mixture regression for observational data, with application to functional regression models. Available at: <http://arxiv.org/abs/1307.0170>.
- JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. and HINTON, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput.* **3** 79–87.
- JIANG, Y., CONGLIAN, Y. and QINGHUA, J. (2018). Model selection for the localized mixture of experts models. *J. Appl. Stat.* **45** 1994–2006. [MR3820067](#) <https://doi.org/10.1080/02664763.2017.1405914>
- KERIBIN, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A* **62** 49–66. [MR1769735](#)
- KHALILI, A. and CHEN, J. (2007). Variable selection in finite mixture of regression models. *J. Amer. Statist. Assoc.* **102** 1025–1038. [MR2411662](#) <https://doi.org/10.1198/016214507000000590>
- KHALILI, A. and LIN, S. (2013). Regularization in finite mixture of regression models with diverging number of parameters. *Biometrics* **69** 436–446. [MR3071062](#) <https://doi.org/10.1111/biom.12020>
- LLOYD-JONES, L. R., NGUYEN, H. D. and MCLACHLAN, G. J. (2018). A globally convergent algorithm for lasso-penalized mixture of linear regression models. *Comput. Statist. Data Anal.* **119** 19–38. [MR3729520](#) <https://doi.org/10.1016/j.csda.2017.09.003>
- MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)* 281–297. Univ. California Press, Berkeley, CA. [MR0214227](#)
- MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models. Wiley Series in Probability and Statistics: Applied Probability and Statistics*. Wiley Interscience, New York. [MR1789474](#) <https://doi.org/10.1002/0471721182>
- MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. [MR1243503](#) <https://doi.org/10.1093/biomet/80.2.267>
- MISITI, M., MISITI, Y., POGGI, J.-M. and PORTIER, B. (2015). Mixture of linear regression models for short term pm10 forecasting in haute normandie (France). *CS-BIGS* **6** 47–60.
- MORVAN, M., DEVIJVER, E., GIACOFCI, M. and MONBET, V. (2021). Supplement to “Prediction of the NASH through penalized mixture of logistic regression models.” <https://doi.org/10.1214/20-AOAS1409SUPP>
- ROSS, J. and DY, J. (2013). Nonparametric mixture of Gaussian processes with constraints. In *Proc. 30th Int. Conf. Mach. Learn.* **28** 1346–1354.
- ROSSET, S. and TIBSHIRANI, R. J. (2020). From Fixed-X to Random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *J. Amer. Statist. Assoc.* **115** 138–151. [MR4078450](#) <https://doi.org/10.1080/01621459.2018.1424632>
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. [MR0468014](#)
- SHMUELI, G. (2010). To explain or to predict? *Statist. Sci.* **25** 289–310. [MR2791669](#) <https://doi.org/10.1214/10-STS330>
- STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ_1 -penalization for mixture regression models. *TEST* **19** 209–256. [MR2677722](#) <https://doi.org/10.1007/s11749-010-0197-z>
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- VINH, N. X., EPPS, J. and BAILEY, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11** 2837–2854. [MR2738784](#)
- WANG, H., LI, R. and TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. [MR2410008](#) <https://doi.org/10.1093/biomet/asm053>
- XU, L., JORDAN, M. I. and HINTON, G. E. (1995). An alternative model for mixtures of experts. In *Advances in Neural Information Processing Systems* 633–640.
- YOUNOSSI, Z., ANSTEE, Q. M., MARIETTI, M., HARDY, T., HENRY, L., ESLAM, M., GEORGE, J. and BUGIANESI, E. (2018a). Global burden of nash and nash: Trends, predictions, risk factors and prevention. *Nature Reviews Gastroenterology & Hepatology* **15** 11–20.
- YOUNOSSI, Z., LOOMBA, R., ANSTEE, Q., RINELLA, M., BUGIANESI, E., MARCHESINI, G., NEUSCHWANDER-TETRI, B., SERFATY, L., NEGRO, F. et al. (2018b). Diagnostic modalities for nonalcoholic fatty liver disease, nonalcoholic steatohepatitis, and associated fibrosis. *Hepatology* **68** 349–360.
- YUKSEL, S. E., WILSON, J. N. and GADER, P. D. (2012). Twenty years of mixture of experts. *IEEE Trans. Neural Netw. Learn. Syst.* **23** 1177–1193. <https://doi.org/10.1109/TNNLS.2012.2200299>

EFFICIENT BAYESIAN INFERENCE OF GENERAL GAUSSIAN MODELS ON LARGE PHYLOGENETIC TREES

BY PAUL BASTIDE¹, LAM SI TUNG HO², GUY BAELE^{3,*}, PHILIPPE LEMEY^{3,†} AND
MARC A. SUCHARD⁴

¹IMAG, CNRS, Université de Montpellier, paul.bastide@umontpellier.fr

²Department of Mathematics and Statistics, Dalhousie University, lam.ho@dal.ca

³Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, *guy.baele@kuleuven.be;
†philippe.lemey@kuleuven.be

⁴Departments of Biostatistics, Biomathematics, and Human Genetics, University of California, Los Angeles,
msuchard@ucla.edu

Phylogenetic comparative methods correct for shared evolutionary history among a set of nonindependent organisms by modeling sample traits as arising from a diffusion process along the branches of a possibly unknown history. To incorporate such uncertainty, we present a scalable Bayesian inference framework under a general Gaussian trait evolution model that exploits Hamiltonian Monte Carlo (HMC). HMC enables efficient sampling of the constrained model parameters and takes advantage of the tree structure for fast likelihood and gradient computations, yielding algorithmic complexity linear in the number of observations. This approach encompasses a wide family of stochastic processes, including the general Ornstein–Uhlenbeck (OU) process, with possible missing data and measurement errors. We implement inference tools for a biologically relevant subset of all these models into the BEAST phylogenetic software package and develop model comparison through marginal likelihood estimation. We apply our approach to study the morphological evolution in the superfamily of *Musteloidea* (including weasels and allies) as well as the heritability of HIV virulence. This second problem furnishes a new measure of evolutionary heritability that demonstrates its utility through a targeted simulation study.

REFERENCES

- ABELES, P. (2016). Efficient Java Matrix Library v0.30.
- AKAIKE, H. (1974). A new look at the statistical model identification: System identification and time-series analysis. *IEEE Trans. Automat. Control* **AC-19** 716–723. [MR0423716 https://doi.org/10.1109/tac.1974.1100705](https://doi.org/10.1109/tac.1974.1100705)
- AL-MOHY, A. H. and HIGHAM, N. J. (2010). The complex step approximation to the Fréchet derivative of a matrix function. *Numer. Algorithms* **53** 113–148. [MR2566131 https://doi.org/10.1007/s11075-009-9323-y](https://doi.org/10.1007/s11075-009-9323-y)
- ALIZON, S., VON WYL, V., STADLER, T., KOUYOS, R. D., YERLY, S., HIRSCHL, B., BÖNI, J., SHAH, C., KLIMKAIT, T. et al. (2010). Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathog.* **6** e1001123. <https://doi.org/10.1371/journal.ppat.1001123>
- ÁLVAREZ-CARRETERO, S., GOSWAMI, A., YANG, Z. and DOS REIS, M. (2019). Bayesian estimation of species divergence times using correlated quantitative characters. *Syst. Biol.* **68** 967–986. <https://doi.org/10.1093/sysbio/syz015>
- ARISTIDE, L. and MORLON, H. (2019). Understanding the effect of competition during evolutionary radiations: An integrated model of phenotypic and species diversification. *Ecol. Lett.* **22** 2006–2017. <https://doi.org/10.1111/ele.13385>
- ARISTIDE, L., DOS REIS, S. F., MACHADO, A. C., LIMA, I., LOPES, R. T. and PEREZ, S. I. (2016). Brain shape convergence in the adaptive radiation of New World monkeys. *Proc. Natl. Acad. Sci. USA* **113** 2158–2163.
- BAELE, G., LEMEY, P. and SUCHARD, M. A. (2016). Genealogical working distributions for Bayesian model testing with phylogenetic uncertainty. *Syst. Biol.* **65** 250–264.
- BAELE, G., SUCHARD, M. A., RAMBAUT, A. and LEMEY, P. (2017). Emerging concepts of data integration in pathogen phylodynamics. *Syst. Biol.* **66** 47–65.

Key words and phrases. Statistical phylogenetics, phylodynamics, Ornstein–Uhlenbeck process, Bayesian inference, Hamiltonian Monte Carlo, model selection, BEAST, heritability, HIV, Musteloidea, total evidence.

- BARTHA, I., MCLAREN, P. J., BRUMME, C., HARRIGAN, R., TELENTI, A. and FELLAY, J. (2017). Estimating the respective contributions of human and viral genetic variation to HIV control. *PLoS Comput. Biol.* **13** e1005339. <https://doi.org/10.1371/journal.pcbi.1005339>
- BARTOSZEK, K. (2017). Trait evolution with jumps: Illusionary normality. Available at [arXiv:1709.04702](https://arxiv.org/abs/1709.04702).
- BARTOSZEK, K., PIENAAR, J., MOSTAD, P., ANDERSSON, S. and HANSEN, T. F. (2012). A phylogenetic comparative method for studying multivariate adaptation. *J. Theoret. Biol.* **314** 204–215. [MR2984458 https://doi.org/10.1016/j.jtbi.2012.08.005](https://doi.org/10.1016/j.jtbi.2012.08.005)
- BARTOSZEK, K., GLÉMIN, S., KAJ, I. and LASCOUX, M. (2017). Using the Ornstein-Uhlenbeck process to model the evolution of interacting populations. *J. Theoret. Biol.* **429** 35–45. [MR3672925 https://doi.org/10.1016/j.jtbi.2017.06.011](https://doi.org/10.1016/j.jtbi.2017.06.011)
- BASTIDE, P. (2017). Shifted stochastic processes evolving on trees: Application to models of adaptive evolution on phylogenies. PhD thesis, Univ. Paris-Saclay.
- BASTIDE, P., MARIADASSOU, M. and ROBIN, S. (2017). Detection of adaptive shifts on phylogenies by using shifted stochastic processes on a tree. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1067–1093. [MR3689309 https://doi.org/10.1111/rssb.12206](https://doi.org/10.1111/rssb.12206)
- BASTIDE, P., ANÉ, C., ROBIN, S. and MARIADASSOU, M. (2018). Inference of adaptive shifts for multivariate correlated traits. *Syst. Biol.* **67** 662–680. <https://doi.org/10.1093/sysbio/syy005>
- BASTIDE, P., HO, L. S., BAELE, G., LEMEY, P. and SUCHARD, M. A. (2021a). Supplement to “Efficient Bayesian inference of general Gaussian models on large phylogenetic trees.” <https://doi.org/10.1214/20-AOAS1419SUPPA>.
- BASTIDE, P., HO, L. S., BAELE, G., LEMEY, P. and SUCHARD, M. A. (2021b). Scripts and data for “Efficient Bayesian inference of general Gaussian models on large phylogenetic trees.” <https://doi.org/10.1214/20-AOAS1419SUPPB>.
- BERGER, S. A. and STAMATAKIS, A. (2010). Accuracy of morphology-based phylogenetic fossil placement under maximum likelihood. In *ACS/IEEE International Conference on Computer Systems and Applications—AICCSA 2010 IEEE*, New York.
- BERTELS, F., MARZEL, A., LEVENTHAL, G., MITOV, V., FELLAY, J., GÜNTARD, H. F., BÖNI, J., YERLY, S., KLIMKAIT, T. et al. (2018). Dissecting HIV virulence: Heritability of setpoint viral load, CD4+ T-cell decline, and per-parasite pathogenicity. *Mol. Biol. Evol.* **35** 27–37. <https://doi.org/10.1093/molbev/msx246>
- BETANCOURT, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. Available at [arXiv:1701.02434](https://arxiv.org/abs/1701.02434).
- BLANQUART, F., WYMANT, C., CORNELISSEN, M., GALL, A., BAKKER, M., BEZEMER, D., HALL, M., HILLEBREGT, M., ONG, S. H. et al. (2017). Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe. *PLoS Biol.* **15** 1–26.
- BOUCHER, F. C., DÉMERY, V., CONTI, E., HARMON, L. J. and UYEDA, J. (2018). A general model for estimating macroevolutionary landscapes. *Syst. Biol.* **67** 304–319. <https://doi.org/10.1093/sysbio/syx075>
- CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in Hidden Markov Models. Springer Series in Statistics*. Springer, New York. [MR2159833](https://doi.org/10.1007/978-1-4939-9966-2)
- CAVALLI-SFORZA, L. L. and EDWARDS, A. W. F. (1967). Phylogenetic analysis: Models and estimation procedures. *Evolution* **21** 550.
- CLAVEL, J., ARISTIDE, L. and MORLON, H. (2019). A penalized likelihood framework for high-dimensional phylogenetic comparative methods and an application to New-World monkeys brain evolution. *Syst. Biol.* **68** 93–116.
- CLAVEL, J., ESCARGUEL, G. and MERCERON, G. (2015). mvmorph: An R package for fitting multivariate evolutionary models to morphometric data. *Methods Ecol. Evol.* **6** 1311–1319.
- COOPER, N., THOMAS, G. H., VENDITTI, C., MEADE, A. and FRECKLETON, R. P. (2016). A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biol. J. Linn. Soc.* **118** 64–77.
- CYBIS, G. B., SINSHEIMER, J. S., BEDFORD, T., MATHER, A. E., LEMEY, P. and SUCHARD, M. A. (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Ann. Appl. Stat.* **9** 969–991. [MR3371344 https://doi.org/10.1214/15-AOAS821](https://doi.org/10.1214/15-AOAS821)
- DRUMMOND, A. J., HO, S. Y. W., PHILLIPS, M. J. and RAMBAUT, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4** e88.
- DRURY, J., CLAVEL, J., MANCEAU, M. and MORLON, H. (2016). Estimating the effect of competition on trait evolution using maximum likelihood inference. *Syst. Biol.* **65** 700–710.
- DRURY, J. P., GREYER, G. F., GARLAND, T. and MORLON, H. (2018). An assessment of phylogenetic tools for analyzing the interplay between interspecific interactions and phenotypic evolution. *Syst. Biol.* **67** 413–427. <https://doi.org/10.1093/sysbio/syx079>
- DUCHEN, P., LEUENBERGER, C., SZILÁGYI, S. M., HARMON, L. J., EASTMAN, J. M., SCHWEIZER, M. and WEGMANN, D. (2017). Inference of evolutionary jumps in large phylogenies using Lévy processes. *Syst. Biol.* **66** 950–963.

- DUCHEN, P., HAUTPHENNE, S., LEHMANN, L. and SALAMIN, N. (2020). Linking micro and macroevolution in the presence of migration. *J. Theoret. Biol.* **486** 110087, 12. [MR4037604 https://doi.org/10.1016/j.jtbi.2019.110087](https://doi.org/10.1016/j.jtbi.2019.110087)
- DUCHEN, P., ALFARO, M. L., ROLLAND, J., SALAMIN, N. and SILVESTRO, D. (2021). On the effect of asymmetrical trait inheritance on models of trait evolution. *Syst. Biol.* **70** 376–388. <https://doi.org/10.1093/sysbio/syaa055>
- DUDAS, G., CARVALHO, L. M., BEDFORD, T., TATEM, A. J., BAELE, G., FARIA, N. R., PARK, D. J., LADNER, J. T., ARIAS, A. et al. (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544** 309–315.
- FAN, Y., WU, R., CHEN, M. H., KUO, L. and LEWIS, P. O. (2011). Choosing among partition models in Bayesian phylogenetics. *Mol. Biol. Evol.* **28** 523–532.
- FELSENSTEIN, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.* **25** 471–492.
- FELSENSTEIN, J. (1985). Phylogenies and the comparative method. *Amer. Nat.* **125** 1–15.
- FELSENSTEIN, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.
- FISHER, A. A., JI, X., ZHANG, Z., LEMEY, P. and SUCHARD, M. A. (2021). Relaxed random walks at scale. *Syst. Biol.* **70** 258–267. <https://doi.org/10.1093/sysbio/syaa056>
- FITZJOHN, R. G. (2010). Quantitative traits and diversification. *Syst. Biol.* **59** 619–633. <https://doi.org/10.1093/sysbio/syq053>
- FITZJOHN, R. G. (2012). Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods Ecol. Evol.* **3** 1084–1092. <https://doi.org/10.1111/j.2041-210X.2012.00234.x>
- FITZJOHN, R. G., MADDISON, W. P. and OTTO, S. P. (2009). Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Syst. Biol.* **58** 595–611. <https://doi.org/10.1093/sysbio/syp067>
- FOURMENT, M., MAGEE, A. F., WHIDDEN, C., BILGE, A., MATSEN, F. A. and MININ, V. N. (2020). 19 dubious ways to compute the marginal likelihood of a phylogenetic tree topology. *Syst. Biol.* **69** 209–220. <https://doi.org/10.1093/sysbio/syz046>
- FRECKLETON, R. P. (2012). Fast likelihood calculations for comparative analyses. *Methods Ecol. Evol.* **3** 940–947.
- GAVRYUSHKINA, A., HEATH, T. A., KSEPKA, D. T., STADLER, T., WELCH, D. and DRUMMOND, A. J. (2017). Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Syst. Biol.* **66** 57–73. <https://doi.org/10.1093/sysbio/syw060>
- GILL, M. S., TUNG HO, L. S., BAELE, G., LEMEY, P. and SUCHARD, M. A. (2016). A relaxed directional random walk model for phylogenetic trait evolution. *Syst. Biol.* **66** syw093.
- GIROLAMI, M. and CALDERHEAD, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 123–214. [MR2814492 https://doi.org/10.1111/j.1467-9868.2010.00765.x](https://doi.org/10.1111/j.1467-9868.2010.00765.x)
- GOLDBERG, E. E., LANCASTER, L. T. and REE, R. H. (2011). Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. *Syst. Biol.* **60** 451–465.
- GOOLSBY, E. W. (2016). Likelihood-based parameter estimation for high-dimensional phylogenetic comparative models: Overcoming the limitations of “distance-based” methods. *Syst. Biol.* **65** 852–870.
- GOOLSBY, E. W., BRUGGEMAN, J. and ANÉ, C. (2017). Rphylopars: Fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods Ecol. Evol.* **8** 22–27.
- HADFIELD, J. D. and NAKAGAWA, S. (2010). General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *J. Evol. Biol.* **23** 494–508. <https://doi.org/10.1111/j.1420-9101.2009.01915.x>
- HANSEN, T. F. (1997). Stabilizing selection and the comparative analysis of adaptation. *Evolution* **51** 1341–1351. <https://doi.org/10.1111/j.1558-5646.1997.tb01457.x>
- HANSEN, T. F. and MARTINS, E. P. (1996). Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution* **50** 1404–1417. <https://doi.org/10.1111/j.1558-5646.1996.tb03914.x>
- HARMON, L. J. (2019). Phylogenetic comparative methods, Version 1, ed. CC-BY-4.0.
- HASEGAWA, M., KISHINO, H. and YANO, T.-A. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22** 160–174.
- HASSLER, G., TOLKOFF, M. R., ALLEN, W. L., HO, L. S. T., LEMEY, P. and SUCHARD, M. A. (2020). Inferring phenotypic trait evolution on large trees with many incomplete measurements. *J. Amer. Statist. Assoc.* 1–39.
- HODCROFT, E., HADFIELD, J. D., FEARNHILL, E., PHILLIPS, A., DUNN, D., O’SHEA, S., PILLAY, D., LEIGH BROWN, A. J., CELIA, A. et al. (2014). The contribution of viral genotype to plasma viral set-point in HIV infection. *PLoS Pathog.* **10**. <https://doi.org/10.1371/journal.ppat.1004112>

- HOFFMAN, M. D. and GELMAN, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15** 1593–1623. [MR3214779](#)
- HOUSWORTH, E. A., MARTINS, E. P. and LYNCH, M. (2004). The phylogenetic mixed model. *Amer. Nat.* **163** 84–96. <https://doi.org/10.1086/380570>
- JEFFREYS, H. (1935). Some tests of significance, treated by the theory of probability. *Math. Proc. Cambridge Philos. Soc.* **31** 203–222.
- JETZ, W., THOMAS, G. H., JOY, J. B., HARTMANN, K. and MOOERS, A. O. (2012). The global diversity of birds in space and time. *Nature* **491** 444–448.
- JI, X., ZHANG, Z., HOLBROOK, A., NISHIMURA, A., BAELE, G., RAMBAUT, A., LEMEY, P. and SUCHARD, M. A. (2020). Gradients do grow on trees: A linear-time $O(N)$ -dimensional gradient for statistical phylogenetics. *Mol. Biol. Evol.* **37** 3047–3060. <https://doi.org/10.1093/molbev/msaa130>
- JONES, A. T., NGUYEN, H. D. and MCLACHLAN, G. J. (2018). logKDE: Log-transformed kernel density estimation. *J. Open Sour. Softw.* **3** 870.
- KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. [MR3363402](#) <https://doi.org/10.1080/01621459.1995.10476572>
- KLOPFSTEIN, S. and SPASOJEVIC, T. (2019). Illustrating phylogenetic placement of fossils using RoguePlots: An example from ichneumonid parasitoid wasps (Hymenoptera, Ichneumonidae) and an extensive morphological matrix. *PLoS ONE* **14** e0212942.
- KOSTIKOVA, A., SILVESTRO, D., PEARMAN, P. B. and SALAMIN, N. (2016). Bridging inter- and intraspecific trait evolution with a hierarchical Bayesian approach. *Syst. Biol.* **65** 417–431.
- LANDIS, M. J., SCHRAIBER, J. G. and LIANG, M. (2013). Phylogenetic analysis using Lévy processes: Finding jumps in the evolution of continuous traits. *Syst. Biol.* **62** 193–204. <https://doi.org/10.1093/sysbio/sys086>
- LARTILLOT, N. (2014). A phylogenetic Kalman filter for ancestral trait reconstruction using molecular data. *Bioinformatics* **30** 488–496.
- LAW, C. J., SLATER, G. J. and MEHTA, R. S. (2018). Lineage diversity and size disparity in musteloidea: Testing patterns of adaptive radiation using molecular and fossil-based methods. *Syst. Biol.* **67** 127–144. <https://doi.org/10.1093/sysbio/syx047>
- LE GALL, J.-F. (2006). Intégration, Probabilités et Processus Aléatoires, Technical Report. Ecole Normale Supérieure, Paris.
- LEBARBIER, É. and MARY-HUARD, T. (2006). Une introduction au critère BIC: Fondements théoriques et interprétation. *J. Soc. Fr. Stat.* **147** 39–57. [MR2500590](#)
- LEMEY, P., RAMBAUT, A., WELCH, J. J. and SUCHARD, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27** 1877–1885.
- LEONARDI, M., LIBRADO, P., DER SARKISSIAN, C., SCHUBERT, M., ALFARHAN, A. H., ALQURAISHI, S. A., AL-RASHEID, K. A. S., GAMBA, C., WILLERSLEV, E. et al. (2016). Evolutionary patterns and processes: Lessons from ancient DNA. *Syst. Biol.* **66** syw059.
- LEVENTHAL, G. E. and BONHOEFFER, S. (2016). Potential pitfalls in estimating viral load heritability. *Trends Microbiol* **24** 687–698. <https://doi.org/10.1016/j.tim.2016.04.008>
- LEWANDOWSKI, D., KUROWICKA, D. and JOE, H. (2009). Generating random correlation matrices based on vines and extended onion method. *J. Multivariate Anal.* **100** 1989–2001. [MR2543081](#) <https://doi.org/10.1016/j.jmva.2009.04.008>
- LYNCH, M. (1991). Methods for the analysis of comparative data in evolutionary biology. *Evolution* **45** 1065–1080. <https://doi.org/10.1111/j.1558-5646.1991.tb04375.x>
- MADDISON, W. P., MIDFORD, P. E. and OTTO, S. P. (2007). Estimating a binary character’s effect on speciation and extinction. *Syst. Biol.* **56** 701–710.
- MAGNUS, J. R. and NEUDECKER, H. (1986). Symmetry, 0-1 matrices and Jacobians: A review. *Econometric Theory* **2** 157–190.
- MANCEAU, M., LAMBERT, A. and MORLON, H. (2016). A unifying comparative phylogenetic framework including traits coevolving across interacting lineages. *Syst. Biol.* **66** syw115.
- MITOV, V., BARTOSZEK, K. and STADLER, T. (2019). Automatic generation of evolutionary hypotheses using mixed Gaussian phylogenetic models. *Proc. Natl. Acad. Sci. USA* 201813823.
- MITOV, V. and STADLER, T. (2018). A practical guide to estimating the heritability of pathogen traits. *Mol. Biol. Evol.* **35** 756–772. <https://doi.org/10.1093/molbev/msx328>
- MITOV, V., BARTOSZEK, K., ASIMOMITIS, G. and STADLER, T. (2020). Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts. *Theor. Popul. Biol.* **131** 66–78. <https://doi.org/10.1016/j.tpb.2019.11.005>
- MÜLLER, N. F., RASMUSSEN, D. A. and STADLER, T. (2017). The structured coalescent and its approximations. *Mol. Biol. Evol.* **34** 2970–2981. <https://doi.org/10.1093/molbev/msx186>

- NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* A. Gelman, G. L. Jones and X. L. Meng, eds.). *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 113–162. CRC Press, Boca Raton, FL. [MR2858447](#)
- NUISMER, S. L. and HARMON, L. J. (2015). Predicting rates of interspecific interaction from phylogenetic trees. *Ecol. Lett.* **18** 17–27. <https://doi.org/10.1111/ele.12384>
- OAKS, J. R., COBB, K. A., MININ, V. N. and LEACHÉ, A. D. (2019). Marginal likelihoods in phylogenetics: A review of methods and applications. *Syst. Biol.* **68** 681–697.
- PAGEL, M. (1999). Inferring the historical patterns of biological evolution. *Nature* **401** 877–884.
- PARINS-FUKUCHI, C. (2018a). Use of continuous traits can improve morphological phylogenetics. *Syst. Biol.* **67** 328–339. <https://doi.org/10.1093/sysbio/syx072>
- PARINS-FUKUCHI, C. (2018b). Bayesian placement of fossils on phylogenies using quantitative morphometric data. *Evolution* **72** 1801–1814. <https://doi.org/10.1111/evo.13516>
- PENNELL, M. W. and HARMON, L. J. (2013). An integrative view of phylogenetic comparative methods: Connections to population genetics, community ecology, and paleobiology. *Ann. N.Y. Acad. Sci.* **1289** 90–105. <https://doi.org/10.1111/nyas.12157>
- PENNELL, M. W., FITZJOHN, R. G., CORNWELL, W. K. and HARMON, L. J. (2015). Model adequacy and the macroevolution of angiosperm functional traits. *Amer. Nat.* **186** E33–E50. <https://doi.org/10.1086/682022>
- PYBUS, O. G., SUCHARD, M. A., LEMEY, P., BERNARDIN, F. J., RAMBAUT, A., CRAWFORD, F. W., GRAY, R. R., ARINAMINPATHY, N., STRAMER, S. L. et al. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci. USA* **109** 15066–15071.
- R CORE TEAM (2019). R: A Language and Environment for Statistical Computing.
- RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77** 257–286.
- RAMBAUT, A., DRUMMOND, A. J., XIE, D., BAELE, G. and SUCHARD, M. A. (2018). Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst. Biol.* **67** 901–904.
- RAZ, R. (2003). On the complexity of matrix product. *SIAM J. Comput.* **32** 1356–1369. [MR2001277](#) <https://doi.org/10.1137/S0097539702402147>
- REVELL, L. J., MAHLER, D. L., REYNOLDS, R. G. and SLATER, G. J. (2015). Placing cryptic, recently extinct, or hypothesized taxa into an ultrametric phylogeny using continuous character data: A case study with the lizard *Anolis roosevelti*. *Evolution* **69** 1027–1035. <https://doi.org/10.1111/evo.12628>
- RONQUIST, F., KLOPFSTEIN, S., VILHELMSSEN, L., SCHULMEISTER, S., MURRAY, D. L. and RASNITSYN, A. P. (2012). A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera. *Syst. Biol.* **61** 973–999.
- SALESA, M. J., ANTÓN, M., SILICEO, G., PESQUERO, M. D., MORALES, J. and ALCALÁ, L. (2013). A non-aquatic otter (Mammalia, Carnivora, Mustelidae) from the Late Miocene (Vallesian, MN 10) of La Roma 2 (Alfambra, Teruel, Spain): Systematics and functional anatomy. *Zool. J. Linn. Soc.* **169** 448–482.
- SCHNITZLER, J., THEIS, C., POLLY, P. D. and ERONEN, J. T. (2017). Fossils matter—understanding modes and rates of trait evolution in Musteloidea (Carnivora). *Evolutionary Ecology Research* **18** 187–200.
- STAN, D. T. (2017). Stan Modeling Language: User’s Guide and Reference Manual v 2.17.0.
- SUCHARD, M. A., LEMEY, P., BAELE, G., AYRES, D. L., DRUMMOND, A. J. and RAMBAUT, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* **4** 1–5.
- TAVARÉ, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17** 57–86.
- TOLKOFF, M. R., ALFARO, M. E., BAELE, G., LEMEY, P. and SUCHARD, M. A. (2018). Phylogenetic factor analysis. *Syst. Biol.* **67** 384–399. <https://doi.org/10.1093/sysbio/syx066>
- UYEDA, J. C. and HARMON, L. J. (2014). A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Syst. Biol.* **63** 902–918.
- VARÓN-GONZÁLEZ, C., WHELAN, S. and KLINGENBERG, C. P. (2020). Estimating phylogenies from shape and similar multidimensional data: Why it is not reliable. *Syst. Biol.* **69** 863–883. <https://doi.org/10.1093/sysbio/syaa003>
- VRANCKEN, B., LEMEY, P., RAMBAUT, A., BEDFORD, T., LONGDON, B., GÜNTARD, H. F. and SUCHARD, M. A. (2015). Simultaneously estimating evolutionary history and repeated traits phylogenetic signal: Applications to viral and host phenotypic evolution. *Methods Ecol. Evol.* **6** 67–82. <https://doi.org/10.1111/2041-210X.12293>
- WANG, L.-G., LAM, T. T.-Y., XU, S., DAI, Z., ZHOU, L., FENG, T., GUO, P., DUNN, C. W., JONES, B. R. et al. (2019). Treeio: An R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.* **37** 599–602. <https://doi.org/10.1093/molbev/msz240>
- WIENS, J. J. (2009). Paleontology, genomics, and combined-data phylogenetics: Can molecular data improve phylogeny estimation for fossil taxa? *Syst. Biol.* **58** 87–99. <https://doi.org/10.1093/sysbio/syp012>

- WIENS, J. J., KUCZYNSKI, C. A., TOWNSEND, T., REEDER, T. W., MULCAHY, D. G. and SITES, J. W. (2010). Combining phylogenomics and fossils in higher-level squamate reptile phylogeny: Molecular data change the placement of fossil taxa. *Syst. Biol.* **59** 674–688.
- XIE, W., LEWIS, P. O., FAN, Y., KUO, L. and CHEN, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **60** 150–160.
- YANG, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution* **39** 306–314.
- YU, G., SMITH, D. K., ZHU, H., GUAN, Y. and LAM, T. T.-Y. (2017). `ggtree`: An `R` package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8** 28–36.
- YU, G., LAM, T. T.-Y., ZHU, H. and GUAN, Y. (2018). Two methods for mapping and visualizing associated data on phylogeny using Ggtree. *Mol. Biol. Evol.* **35** 3041–3043.

A COVARIANCE-ENHANCED APPROACH TO MULTITISSUE JOINT EQTL MAPPING WITH APPLICATION TO TRANSCRIPTOME-WIDE ASSOCIATION STUDIES

BY AARON J. MOLSTAD¹, WEI SUN^{2,*} AND LI HSU^{2,†}

¹Department of Statistics, Genetics Institute, University of Florida, amolstad@ufl.edu

²Biostatistics Program, Fred Hutchinson Cancer Research Center, *wsun@fredhutch.org; †lih@fredhutch.org

Transcriptome-wide association studies based on genetically predicted gene expression have the potential to identify novel regions associated with various complex traits. It has been shown that incorporating expression quantitative trait loci (eQTLs) corresponding to multiple tissue types can improve power for association studies involving complex etiology. In this article we propose a new multivariate response linear regression model and method for predicting gene expression in multiple tissues simultaneously. Unlike existing methods for multitissue joint eQTL mapping, our approach incorporates tissue-tissue expression correlation which allows us to more efficiently handle missing expression measurements and to more accurately predict gene expression using a weighted summation of eQTL genotypes. We show through simulation studies that our approach performs better than the existing methods in many scenarios. We use our method to estimate eQTL weights for 29 tissues collected by GTEx, and show that our approach significantly improves expression prediction accuracy compared to competitors. Using our eQTL weights, we perform a multitissue-based S-MultiXcan (*PLoS Genet.* **15** (2019) e1007889) transcriptome-wide association study and show that our method leads to more discoveries in novel regions and more discoveries overall than the existing methods. Estimated eQTL weights and code for implementing the method are available for download online at github.com/ajmolstad/MTeQTLResults.

REFERENCES

- BARBEIRA, A. N., DICKINSON, S. P., BONAZZOLA, R., ZHENG, J., WHEELER, H. E., TORRES, J. M., TORSTENSON, E. S., SHAH, K. P., GARCIA, T. et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9** 1825. <https://doi.org/10.1038/s41467-018-03621-1>
- BARBEIRA, A. N., PIVIDORI, M. D., ZHENG, J., WHEELER, H. E., NICOLAE, D. L., IM, H. K. et al. (2019). Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* **15** e1007889.
- BUNIELLO, A., MACARTHUR, J. A. L., CEREZO, M., HARRIS, L. W., HAYHURST, J., MALANGONE, C., MCMAHON, A., MORALES, J., MOUNTJOY, E. et al. (2018). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47** D1005–D1012.
- BUNIELLO, A., MACARTHUR, J. A. L., CEREZO, M., HARRIS, L. W., HAYHURST, J., MALANGONE, C., MCMAHON, A., MORALES, J., MOUNTJOY, E. et al. (2019). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47** D1005–D1012.
- CHEN, L., POURAHMADI, M. and MAADOOLAT, M. (2014). Regularized multivariate regression models with skew- t error distributions. *J. Statist. Plann. Inference* **149** 125–139. [MR3199899 https://doi.org/10.1016/j.jspi.2014.02.001](https://doi.org/10.1016/j.jspi.2014.02.001)
- CHEUNG, V. G., CONLIN, L. K., WEBER, T. M., ARCARO, M., JEN, K. Y., MORLEY, M. and SPIELMAN, R. S. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* **33** 422–425.

Key words and phrases. Transcriptome-wide association studies, multivariate regression, expression quantitative trait loci, multitissue integrative analysis, GTEx.

- GTEX CONSORTIUM (2015). The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348** 648–660.
- GTEX CONSORTIUM (2017). Genetic effects on gene expression across human tissues. *Nature* **550** 204.
- FLUTRE, T., WEN, X., PRITCHARD, J. and STEPHENS, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* **9** e1003486. <https://doi.org/10.1371/journal.pgen.1003486>
- GAMAZON, E. R., WHEELER, H. E., SHAH, K. P., MOZAFFARI, S. V., AQUINO-MICHAELS, K., CARROLL, R. J., EYLER, A. E., DENNY, J. C., GTEX CONSORTIUM et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47** 1091–1098. <https://doi.org/10.1038/ng.3367>
- GUSEV, A., KO, A., SHI, H., BHATIA, G., CHUNG, W., PENNINX, B. W., JANSEN, R., DE GEUS, E. J. et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48** 245–252.
- HINDORFF, L. A., SETHUPATHY, P., JUNKINS, H. A., RAMOS, E. M., MEHTA, J. P., COLLINS, F. S. and MANOLIO, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106** 9362–9367.
- HSIEH, C.-J., SUSTIK, M. A., DHILLON, I. S. and RAVIKUMAR, P. (2014). QUIC: Quadratic approximation for sparse inverse covariance estimation. *J. Mach. Learn. Res.* **15** 2911–2947. [MR3277149](https://doi.org/10.1137/12M0937714)
- HU, Y., LI, M., LU, Q., WENG, H., WANG, J., ZEKAVAT, S. M., YU, Z., LI, B., GU, J. et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet.* **51** 568–576. <https://doi.org/10.1038/s41588-019-0345-7>
- KICHAEV, G., BHATIA, G., LOH, P.-R., GAZAL, S., BURCH, K., FREUND, M. K., SCHOECH, A., PASANIUC, B. and PRICE, A. L. (2019). Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* **104** 65–75. <https://doi.org/10.1016/j.ajhg.2018.11.008>
- LALONDE, S., CODINA-FAUTEUX, V.-A., DE BELLEFON, S. M., LEBLANC, F., BEAUDOIN, M., SIMON, M.-M., DALI, R., KWAN, T., LO, K. S. et al. (2019). Integrative analysis of vascular endothelial cell genomic features identifies AIDA as a coronary artery disease candidate gene. *Genome Biol.* **20** 1–13.
- LANGE, K. (2016). *MM Optimization Algorithms*. SIAM, Philadelphia, PA. [MR3522165 https://doi.org/10.1137/1.9781611974409.ch1](https://doi.org/10.1137/1.9781611974409.ch1)
- LEE, W. and LIU, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *J. Multivariate Anal.* **111** 241–255. [MR2944419 https://doi.org/10.1016/j.jmva.2012.03.013](https://doi.org/10.1016/j.jmva.2012.03.013)
- LI, G., SHABALIN, A. A., RUSYN, I., WRIGHT, F. A. and NOBEL, A. B. (2018). An empirical Bayes approach for multiple tissue eQTL analysis. *Biostatistics* **19** 391–406. [MR3815178 https://doi.org/10.1093/biostatistics/kxx048](https://doi.org/10.1093/biostatistics/kxx048)
- MANCUSO, N., SHI, H., GODDARD, P., KICHAEV, G., GUSEV, A. and PASANIUC, B. (2017). Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.* **100** 473–487. <https://doi.org/10.1016/j.ajhg.2017.01.031>
- MANOR, O. and SEGAL, E. (2015). GenoExp: A web tool for predicting gene expression levels from single nucleotide polymorphisms. *Bioinformatics* **31** 1848–1850.
- MAURANO, M. T., HUMBERT, R., RYNES, E., THURMAN, R. E., HAUGEN, E., WANG, H., REYNOLDS, A. P., SANDSTROM, R. et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337** 1190–1195.
- MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. [MR1243503 https://doi.org/10.1093/biomet/80.2.267](https://doi.org/10.1093/biomet/80.2.267)
- MOLSTAD, A. J. and ROTHMAN, A. J. (2018). Shrinking characteristics of precision matrix estimators. *Biometrika* **105** 563–574. [MR3842885 https://doi.org/10.1093/biomet/asy023](https://doi.org/10.1093/biomet/asy023)
- MOLSTAD, A. J., SUN, W. and HSU, L. (2021). Supplement to “A covariance-enhanced approach to multitissue joint eQTL mapping with application to transcriptome-wide association studies.” <https://doi.org/10.1214/20-AOAS1432SUPP>
- MONTGOMERY, S. B., SAMMETH, M., GUTIERREZ-ARCELUS, M., LACH, R. P., INGLE, C., NISBETT, J., GUIGO, R. and DERMITZAKIS, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464** 773–777.
- MORGAN, A., VUCKOVIC, D., KRISHNAMOORTHY, N., RUBINATO, E., AMBROSETTI, U., CASTORINA, P., FRANZÈ, A., VOZZI, D., BIANCA, M. L. et al. (2019). Next-generation sequencing identified SPATC1L as a possible candidate gene for both early-onset and age-related hearing loss. *Eur. J. Hum. Genet.* **27** 70–79. <https://doi.org/10.1038/s41431-018-0229-9>
- NEALE, B. (2019). UKBiobank GWAS summary statistics.
- NICOLAE, D. L., GAMAZON, E., ZHANG, W., DUAN, S., DOLAN, M. E. and COX, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genet.* **6** e1000888. <https://doi.org/10.1371/journal.pgen.1000888>

- PARIKH, N. and BOYD, S. (2014). Proximal algorithms. *Found. Trends Optim.* **1** 127–239.
- PAVLIDES, J. M., ZHU, Z., GRATTEN, J., MCRAE, A. F., WRAY, N. R. and YANG, J. (2016). Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. *Gen. Med.* **8** 84.
- PENG, J., ZHU, J., BERGAMASCHI, A., HAN, W., NOH, D.-Y., POLLACK, J. R. and WANG, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.* **4** 53–77. [MR2758084](#) <https://doi.org/10.1214/09-AOAS271>
- PICKRELL, J. K., MARIONI, J. C., PAI, A. A., DEGNER, J. F., ENGELHARDT, B. E., NKADORI, E., VEYRIERAS, J. B., STEPHENS, M. et al. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464** 768–772.
- PILLING, L. C., ATKINS, J. L., DUFF, M. O., BEAUMONT, R. N., JONES, S. E., TYRRELL, J., KUO, C.-L., RUTH, K. S., TUKE, M. A. et al. (2017). Red blood cell distribution width: Genetic evidence for aging pathways in 116,666 volunteers. *PLoS ONE* **12** e0185083. <https://doi.org/10.1371/journal.pone.0185083>
- POLSON, N. G., SCOTT, J. G. and WILLARD, B. T. (2015). Proximal algorithms in statistics and machine learning. *Statist. Sci.* **30** 559–581. [MR3432841](#) <https://doi.org/10.1214/15-STS530>
- ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2010). Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Statist.* **19** 947–962. Supplementary materials available online. [MR2791263](#) <https://doi.org/10.1198/jcgs.2010.09188>
- ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. [MR2417391](#) <https://doi.org/10.1214/08-EJS176>
- SCHADT, E. E., MONKS, S. A., DRAKE, T. A., LUSIS, A. J., CHE, N., COLINAYO, V., RUFF, T. G., MILLIGAN, S. B., LAMB, J. R. et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422** 297–302. <https://doi.org/10.1038/nature01434>
- SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. [MR3173712](#) <https://doi.org/10.1080/10618600.2012.681250>
- SRIVASTAVA, V. K. and GILES, D. E. A. (1987). *Seemingly Unrelated Regression Equations Models: Estimation and Inference. Statistics: Textbooks and Monographs* **80**. Dekker, New York. [MR0930104](#)
- STRANGER, B. E., FORREST, M. S., CLARK, A. G., MINICHELLO, M. J., DEUTSCH, S., LYLE, R., HUNT, S., KAHL, B., ANTONARAKIS, S. E. et al. (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1** e78.
- TSELIU, E., TERROVITIS, J. V., KALDARA, E. E., NTALIANIS, A. S., REPASOS, E., KATSAROS, L., MARGARI, Z. J., MATSOUKA, C., TOUMANIDIS, S. et al. (2014). Red blood cell distribution width is a significant prognostic marker in advanced heart failure, independent of hemoglobin levels. *Ell. Kardiol. Epiteor.* **55** 457–461.
- VISSCHER, P. M., HILL, W. G. and WRAY, N. R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9** 255–266.
- WANG, J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Statist. Sinica* **25** 831–851. [MR3409726](#)
- WANG, J., GAMAZON, E. R., PIERCE, B. L., STRANGER, B. E., IM, H. K., GIBBONS, R. D., COX, N. J., NICOLAE, D. L. and CHEN, L. S. (2016). Imputing gene expression in uncollected tissues within and beyond GTEx. *Am. J. Hum. Genet.* **98** 697–708.
- XANTHOPOULOS, A., GIAMOUZIS, G., MELIDONIS, A., KITAI, T., PARASKEVOPOULOU, E., PARASKEVOPOULOU, P., PATSILINAKOS, S., TRIPOSKIADIS, F. and SKOULARIGIS, J. (2017). Red blood cell distribution width as a prognostic marker in patients with heart failure and diabetes mellitus. *Cardiovasc. Diabetol.* **16** 81. <https://doi.org/10.1186/s12933-017-0563-1>
- YIN, J. and LI, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5** 2630–2650. [MR2907129](#) <https://doi.org/10.1214/11-AOAS494>
- YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. [MR2367824](#) <https://doi.org/10.1093/biomet/asm018>
- ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57** 348–368. [MR0139235](#)
- ZENG, P., ZHOU, X. and HUANG, S. (2017). Prediction of gene expression with cis-SNPs using mixed models and regularization methods. *BMC Genomics* **18** 368. <https://doi.org/10.1186/s12864-017-3759-6>
- ZHAO, G., YANG, W., WU, J., CHEN, B., YANG, X., CHEN, J., MCVEY, D. G., ANDREADI, C., GONG, P. et al. (2018). Influence of a coronary artery disease-associated genetic variant on FURIN expression and effect of FURIN on macrophage behavior. *Arterioscler. Thromb. Vasc. Biol.* **38** 1837–1844.

TENSOR-VARIATE FINITE MIXTURE MODELING FOR THE ANALYSIS OF UNIVERSITY PROFESSOR REMUNERATION

BY SHUCHISMITA SARKAR¹, VOLODYMYR MELNYKOV^{2,*} AND XUWEN ZHU^{2,†}

¹Department of Applied Statistics and Operations Research, Bowling Green State University, ssarkar@bgsu.edu

²Department of Information Systems, Statistics, and Management Science, University of Alabama, *vmelnykov@ua.edu;

†xzhu20@cba.ua.edu

There has been a long-standing interest in the analysis of university professor salary data. The vast majority of the publications on the topic employ linear regression models in an attempt to predict individual salaries. Indeed, the administration of any academic institution is interested in adequately compensating the faculty to attract and keep the best specialists available on the market. However, higher administration and legislators are not concerned with the matter of individual compensation and need to have a bigger picture for developing university strategies and policies. This paper is the first attempt to model university compensation data at the institutional level. The analysis of university salary patterns is a challenging problem due to the heterogeneous, skewed, multiway and temporal nature of the data. This paper aims at addressing all the above-mentioned issues by proposing a novel tensor regression mixture model and applying it to the data set obtained from the American Association of University Professors. The utility of the developed model is illustrated on addressing several important questions related to gender equity and peer institution comparison.

REFERENCES

- ASHRAF, J. and SHABIR, T. (2006). Are there racial differences in faculty salaries? *J. Econ. Finance* **30** 306–316.
- BANFIELD, J. D. and RAFTERY, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803–821. [MR1243494 https://doi.org/10.2307/2532201](https://doi.org/10.2307/2532201)
- BASSER, P. J. and PAJEVIC, S. (2003). A normal distribution for tensor-valued random variables: Applications to diffusion tensor MRI. *IEEE Trans. Med. Imag.* **22** 785–794.
- BASU, S., BANERJEE, A. and MOONEY, R. J. (2004). Active semi-supervision for pairwise constrained clustering. In *Proceedings of the Fourth SIAM International Conference on Data Mining* 333–344. SIAM, Philadelphia, PA. [MR2388453](https://doi.org/10.1137/0403016)
- BECKER, G. S. (1975). Front matter, human capital: A theoretical and empirical analysis, with special reference to education. In *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education* 1–22 2nd ed. NBER.
- BECKER, W. E. and TOUTKOUSIAN, R. K. (2003). Measuring gender bias in the salaries of tenured faculty members. *New Directions for Institutional Research* **117** 5–20.
- BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput. Statist. Data Anal.* **413** 561–575. [MR1968069 https://doi.org/10.1016/S0167-9473\(02\)00163-9](https://doi.org/10.1016/S0167-9473(02)00163-9)
- BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. (With discussion). *J. Roy. Statist. Soc. Ser. B* **26** 211–252. [MR0192611](https://doi.org/10.1093/bjbs/26.3.211)
- BROWNE, R. P. and MCNICHOLAS, P. D. (2015). A mixture of generalized hyperbolic distributions. *Canad. J. Statist.* **43** 176–198. [MR3353379 https://doi.org/10.1002/cjs.11246](https://doi.org/10.1002/cjs.11246)
- COHN, E. (1973). Factors affecting variations in faculty salaries and compensation in institutions of higher education. *The Journal of Higher Education* **44** 124–136.
- DELORME, C. D. J., HILL, R. C. and WOOD, N. J. (1979). Analysis of a quantitative method of determining faculty salaries. *Journal of Economic Education* **11** 20–25.
- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38. [MR0501537](https://doi.org/10.1093/bjbs/39.1.1)

- FAIRWEATHER, J. S. (1993). Faculty reward structures: Toward institutional and professional homogenization. *Res. High. Educ.* **34** 603–623.
- FERREIRA, A. P. (2013). Are all faculty members being compensated equally? A multi-method approach to investigating faculty salary. Open Access Dissertation.
- FRANCZAK, B. C., BROWNE, R. P. and MCNICHOLAS, P. D. (2014). Mixtures of shifted asymmetric Laplace distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* **36** 1149–1157.
- FRIEDMAN, M. and KUZNETS, S. (1945). *Income from Independent Professional*. National Bureau of Economic Research, New York.
- GALLAUGHER, M. P. B. and MCNICHOLAS, P. D. (2017). A matrix variate skew- t distribution. *Stat* **6** 160–170. MR3653050 <https://doi.org/10.1002/sta4.143>
- GALLAUGHER, M. P. B. and MCNICHOLAS, P. D. (2018). Finite mixtures of skewed matrix variate distributions. *Pattern Recognit.* **80** 83–93.
- GALLAUGHER, M. P. B. and MCNICHOLAS, P. D. (2019). Package MatSkew: Matrix skew- t parameter estimation.
- GALLAUGHER, M. P. B. and MCNICHOLAS, P. D. (2019c). Three skewed matrix variate distributions. *Statist. Probab. Lett.* **145** 103–109. MR3873895 <https://doi.org/10.1016/j.spl.2018.08.012>
- HEARN, J. C. (1999). Pay and performance in the university: An examination of faculty salaries. *The Review of Higher Education* **22** 391–410.
- HEXTER, H. (1990). Faculty salaries in perspective. *Research Briefs* **1**.
- HILL, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology* **54** 427–432.
- LEE, S. X. and MCLACHLAN, G. J. (2013a). Model-based clustering and classification with non-normal mixture distributions. *Stat. Methods Appl.* **22** 427–454. MR3127088 <https://doi.org/10.1007/s10260-013-0237-4>
- LEE, S. X. and MCLACHLAN, G. J. (2013b). On mixtures of skew normal and skew t -distributions. *Adv. Data Anal. Classif.* **7** 241–266. MR3103965 <https://doi.org/10.1007/s11634-013-0132-8>
- LIN, T.-I., HO, H. J. and LEE, C.-R. (2014). Flexible mixture modelling using the multivariate skew- t -normal distribution. *Stat. Comput.* **24** 531–546. MR3223539 <https://doi.org/10.1007/s11222-013-9386-4>
- LO, K. and GOTTARDO, R. (2012). Flexible mixture modeling via the multivariate t distribution with the Box–Cox transformation: An alternative to the skew- t distribution. *Stat. Comput.* **22** 33–52. MR2865054 <https://doi.org/10.1007/s11222-010-9204-1>
- MANCEUR, A. M. and DUTILLEUL, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *J. Comput. Appl. Math.* **239** 37–49. MR2991957 <https://doi.org/10.1016/j.cam.2012.09.017>
- MANLY, B. F. J. (1976). Exponential data transformations. *Biometrics Unit* **25** 37–42.
- MELGUIZO, T. and STROBER, M. H. (2007). Faculty salaries and the maximization of prestige. *Res. High. Educ.* **48** 633–668.
- MELNYKOV, V., MELNYKOV, I. and MICHAEL, S. (2016). Semi-supervised model-based clustering with positive and negative constraints. *Adv. Data Anal. Classif.* **10** 327–349. MR3541239 <https://doi.org/10.1007/s11634-015-0200-3>
- MELNYKOV, V. and ZHU, X. (2018). On model-based clustering of skewed matrix data. *J. Multivariate Anal.* **167** 181–194. MR3830641 <https://doi.org/10.1016/j.jmva.2018.04.007>
- MELNYKOV, V. and ZHU, X. (2019). Studying crime trends in the USA over the years 2000–2012. *Adv. Data Anal. Classif.* **13** 325–341. MR3935201 <https://doi.org/10.1007/s11634-018-0326-1>
- MINCER, J. (1958). Investment in human capital and personal income distribution. *J. Polit. Econ.* **66** 281–302.
- MOHANTY, D., DODDER, R. and KARMAN, T. (1986). Faculty salary analysis by region, rank, and discipline from 1977–1978 to 1983–1984. *Res. High. Educ.* **24** 304–317.
- O’HAGAN, A., MURPHY, T. B., GORMLEY, I. C., MCNICHOLAS, P. D. and KARLIS, D. (2016). Clustering with the multivariate normal inverse Gaussian distribution. *Comput. Statist. Data Anal.* **93** 18–30. MR3406193 <https://doi.org/10.1016/j.csda.2014.09.006>
- PERNA, L. W. (2001). Sex differences in faculty salaries: A cohort analysis. *The Review of Higher Education* **24** 283–307.
- RIPPNER, J. A. and TOUTKOUSIAN, R. K. (2015). The ‘Big Bang’ in public and private faculty salaries. *Journal of Education Finance* **41** 103–123.
- SARKAR, S., MELNYKOV, V. and ZHU, X. (2021). Supplement to “Tensor-variate finite mixture modeling for the analysis of university professor remuneration.” <https://doi.org/10.1214/20-AOAS1420SUPP>
- SARKAR, S., ZHU, X., MELNYKOV, V. and INGRASSIA, S. (2020). On parsimonious models for modeling matrix data. *Comput. Statist. Data Anal.* **142** 106822, 26. MR3992446 <https://doi.org/10.1016/j.csda.2019.106822>
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014
- SHENTAL, N., BAR-HILLEL, A., HERTZ, T. and WEINSHALL, D. (2003). Computing Gaussian mixture models with EM using equivalence constraints. In *Advances in NIPS* **15**.

- SIMPSON, W. B. (1981). Faculty salary structure for a college or university. *The Journal of Higher Education* **52** 219–236.
- SNYDER, J. K., HYER, P. B. and MCCLAUGHLIN, G. W. (1994). Faculty salary equity: Issues and options. *Res. High. Educ.* **35** 1–19.
- TOUTKOUSHIAN, R. K., BELLAS, M. L. and MOORE, J. V. (2007). The interaction effects of gender, race, and marital status on faculty salaries. *The Journal of Higher Education* **78** 572–601.
- UMBACH, P. D. (2007). Gender equity in the academic labor market: An analysis of academic disciplines. *Res. High. Educ.* **48** 169–192.
- VIROLI, C. (2011a). Finite mixtures of matrix normal distributions for classifying three-way data. *Stat. Comput.* **21** 511–522. [MR2826689 https://doi.org/10.1007/s11222-010-9188-x](https://doi.org/10.1007/s11222-010-9188-x)
- VIROLI, C. (2011b). Model based clustering for three-way data structures. *Bayesian Anal.* **6** 573–602. [MR2869958 https://doi.org/10.1214/11-BA622](https://doi.org/10.1214/11-BA622)
- VIROLI, C. (2012). On matrix-variate regression analysis. *J. Multivariate Anal.* **111** 296–309. [MR2944423 https://doi.org/10.1016/j.jmva.2012.04.005](https://doi.org/10.1016/j.jmva.2012.04.005)
- WHITE, A. and MURPHY, T. B. (2016). Exponential family mixed membership models for soft clustering of multivariate data. *Adv. Data Anal. Classif.* **10** 521–540. [MR3575730 https://doi.org/10.1007/s11634-016-0267-5](https://doi.org/10.1007/s11634-016-0267-5)
- YEO, I.-K. and JOHNSON, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika* **87** 954–959. [MR1813988 https://doi.org/10.1093/biomet/87.4.954](https://doi.org/10.1093/biomet/87.4.954)
- ZHU, X. and MELNYKOV, V. (2018). Manly transformation in finite mixture modeling. *Comput. Statist. Data Anal.* **121** 190–208. [MR3759207 https://doi.org/10.1016/j.csda.2016.01.015](https://doi.org/10.1016/j.csda.2016.01.015)
- ZHU, X. and MELNYKOV, V. (2020). MatTransMix: An R package for clustering matrices. R package version 0.1.9.

INFLECTION POINTS IN COMMUNITY-LEVEL HOMELESS RATES

BY CHRIS GLYNN¹, THOMAS H. BYRNE² AND DENNIS P. CULHANE³

¹Zillow Research, christophergl@zillowgroup.com

²School of Social Work, Boston University, tbyrne@bu.edu

³School of Social Policy & Practice, University of Pennsylvania, culhane@upenn.edu

Statistical models of community-level homeless rates typically assume a linear relationship to covariates. This linear model assumption precludes the possibility of inflection points in homeless rates—thresholds in quantifiable metrics of a community that, once breached, are associated with large increases in homelessness. In this paper we identify points of structural change in the relationship between homeless rates and community-level measures of housing affordability and extreme poverty. We utilize the Ewens–Pitman attraction (EPA) distribution to develop a Bayesian nonparametric regression model in which clusters of communities with similar covariates share common patterns of variation in homeless rates. A main finding of the study is that the expected homeless rate in a community begins to quickly increase once median rental costs exceed 30% of median income, providing a statistical link between homelessness and the U.S. government’s definition of a housing cost burden. Our analysis also identifies clusters of communities that exhibit distinct geographic patterns and yields insight into the homelessness and housing affordability crisis unfolding on both coasts of the United States.

REFERENCES

- ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. [MR0365969](#)
- BLEI, D. M. and FRAZIER, P. I. (2011). Distance dependent Chinese restaurant processes. *J. Mach. Learn. Res.* **12** 2461–2488. [MR2834504](#)
- BUN, Y. (2012). Zillow rent index: Methodology. Available at <https://www.zillow.com/research/zillow-rent-index-methodology-2393/>. Accessed 04/2/2017.
- BYRNE, T. (2018). HUD-CoC-Geography-Crosswalk. Available at <https://github.com/tomhbyrne/HUD-CoC-Geography-Crosswalk>.
- BYRNE, T., MUNLEY, E. A., FARGO, J. D., MONTGOMERY, A. E. and CULHANE, D. P. (2013). New perspectives on community-level determinants of homelessness. *J. Urban Aff.* **35** 607–625.
- CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. [MR2758172](#) <https://doi.org/10.1214/09-AOAS285>
- CORINTH, K. C. (2015). Ending homelessness: More housing or fewer shelters? AEI Economics Working Papers 863788, American Enterprise Institute.
- CULHANE, D. P., LEE, C. and WACHTER, S. M. (1996). Where the homeless come from: A study of the prior address distribution of families admitted to public shelters in New York City and Philadelphia. *Hous. Policy Debate* **7** 327–365.
- DAHL, D. B., DAY, R. and TSAI, J. W. (2017). Random partition distribution indexed by pairwise information. *J. Amer. Statist. Assoc.* **112** 721–732. [MR3671765](#) <https://doi.org/10.1080/01621459.2016.1165103>
- ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. [MR1340510](#)
- FRITSCH, A. and ICKSTADT, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Anal.* **4** 367–391. [MR2507368](#) <https://doi.org/10.1214/09-BA414>
- GLYNN, C., BYRNE, T. H. and CULHANE, D. P. (2021). Supplement to “Inflection points in community-level homeless rates.” <https://doi.org/10.1214/20-AOAS1414SUPP>
- GLYNN, C. and FOX, E. B. (2019). Dynamics of homelessness in urban America. *Ann. Appl. Stat.* **13** 573–605. [MR3937441](#) <https://doi.org/10.1214/18-AOAS1200>

- HOPPER, K., SHINN, M., LASKA, E., MEISNER, M. and WANDERLING, J. (2008). Estimating numbers of unsheltered homeless people through plant-capture and postcount survey methods. *Am. J. Publ. Health* **98** 1438–1442.
- HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.
- HUD (2014). Rental burdens: Rethinking affordability measures. Available at https://www.huduser.gov/portal/pdredge/pdr_edge_featd_article_092214.html.
- HUD (2017). Pit and hic data since 2007. Available at <https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/>. Accessed 08/7/2018.
- LEE, B. A., PRICE-SPRATLEN, T. and KANAN, J. W. (2003). Determinants of homelessness in metropolitan areas. *J. Urban Aff.* **25** 335–356.
- MACEACHERN, S. N. (2000). Dependent dirichlet processes 1–40. Unpublished manuscript, Department of Statistics, The Ohio State Univ.
- MÜLLER, P., QUINTANA, F. and ROSNER, G. L. (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist.* **20** 260–278. MR2816548 <https://doi.org/10.1198/jcgs.2011.09066>
- NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 <https://doi.org/10.2307/1390653>
- PAGE, G. L. and QUINTANA, F. A. (2018). Calibrating covariate informed product partition models. *Stat. Comput.* **28** 1009–1031. MR3835631 <https://doi.org/10.1007/s11222-017-9777-z>
- POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712 <https://doi.org/10.1080/01621459.2013.829001>
- QUIGLEY, J. M., RAPHAEL, S. and SMOLENSKY, E. (2001). Homeless in America, homeless in California. *Rev. Econ. Stat.* **83** 37–51.
- RUKMANA, D. (2008). Where the homeless children and youth come from: A study of the residential origins of the homeless in Miami-Dade County, Florida. *Child. Youth Serv. Rev.* **30** 1009–1021.
- R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27** 1413–1432. MR3647105 <https://doi.org/10.1007/s11222-016-9696-4>
- WEST, M. and HARRISON, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1482232
- WILLIAMS, C. K. and RASMUSSEN, C. E. (1996). Gaussian processes for regression. In *Advances in Neural Information Processing Systems* 514–520.

LENGTH-BIASED SEMICOMPETING RISKS MODELS FOR CROSS-SECTIONAL DATA: AN APPLICATION TO CURRENT DURATION OF PREGNANCY ATTEMPT DATA

BY ALEXANDER C. MCLAIN^{1,*}, SIYUAN GUO^{1,†}, MARIE THOMA^{2,§} AND JIAJIA ZHANG^{1,‡}

¹*Department of Epidemiology and Biostatistics, University of South Carolina, *mclaina@mailbox.sc.edu; †siyuang@email.sc.edu; ‡jzhang@mailbox.sc.edu*

²*Department of Family Health Services, University of Maryland, §mthoma@umd.edu*

Cross-sectional length-biased data arise from questions on the at-risk time for an event of interest from those who are at risk but have yet to experience the event. For example, in the National Survey on Family Growth (NSFG) women, who were currently attempting to become pregnant, were asked how long they had been attempting pregnancy. Cross-sectional survival analysis methods use the observed at-risk times to make inference on the distribution of the unobserved time-to-failure. However, methodological gaps in these methods remain such as how to handle semicompeting risks. For example, if the women attempting pregnancy had undergone fertility treatment during their current pregnancy attempt. In this paper we develop statistical methods that extend cross-sectional survival analysis methods to incorporate semicompeting risks. They can be used to estimate the distribution of the length of natural pregnancy attempts (i.e., without fertility treatment) while correctly accounting for women that sought fertility treatment prior to being sampled using cross-sectional data. We demonstrate our approach based on simulated data and an analysis of data from the NSFG. The proposed method results in separate survival curves for time-to-natural-pregnancy, time-to-fertility treatment and time-to-pregnancy after fertility treatment.

REFERENCES

- ALLISON, P. D. (1985). Survival analysis of backward recurrence times. *J. Amer. Statist. Assoc.* **80** 315–322.
- ANDERSEN, P. K. and KEIDING, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Stat. Med.* **31** 1074–1088. [MR2925679](https://doi.org/10.1002/sim.4385) <https://doi.org/10.1002/sim.4385>
- BERGERON, P.-J., ASGHARIAN, M. and WOLFSON, D. B. (2008). Covariate bias induced by length-biased sampling of failure times. *J. Amer. Statist. Assoc.* **103** 737–742. [MR2524006](https://doi.org/10.1198/016214508000000382) <https://doi.org/10.1198/016214508000000382>
- CARONE, M., ASGHARIAN, M. and JEWELL, N. P. (2014). Estimating the lifetime risk of dementia in the Canadian elderly population using cross-sectional cohort survival data. *J. Amer. Statist. Assoc.* **109** 24–35. [MR3180544](https://doi.org/10.1080/01621459.2013.859076) <https://doi.org/10.1080/01621459.2013.859076>
- CHANDRA, A., MARTINEZ, G. M., MOSHER, W. D., ABMA, J. C. and JONES, J. (2005). Fertility, family planning, and reproductive health of U.S. women: Data from the 2002 National Survey of Family Growth. *Vital Health Stat.* **25** 1–160.
- CHENG, Y.-J. and WANG, M.-C. (2012). Estimating propensity scores and causal survival functions using prevalent survival data. *Biometrics* **68** 707–716. [MR3055175](https://doi.org/10.1111/j.1541-0420.2012.01754.x) <https://doi.org/10.1111/j.1541-0420.2012.01754.x>
- CHENG, Y.-J. and WANG, M.-C. (2015). Causal estimation: Using semiparametric transformation models under prevalent sampling. *Biometrics* **71** 302–312. [MR3366235](https://doi.org/10.1111/biom.12286) <https://doi.org/10.1111/biom.12286>
- COX, D. R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling* (N. L. Johnson and H. Smith, eds.) 506–527. Wiley, New York.
- DAUXOIS, J.-Y., GUILLOUX, A. and KIRMANI, S. N. U. A. (2014). Estimation in a competing risks proportional hazards model under length-biased sampling with censoring. *Lifetime Data Anal.* **20** 276–302. [MR3181015](https://doi.org/10.1007/s10985-013-9248-6) <https://doi.org/10.1007/s10985-013-9248-6>

- DURON, S., SLAMA, R., DUCOT, B., BOHET, A., SØRENSEN, D. N., KEIDING, N., MOREAU, C. and BOUYER, J. (2013). Cumulative incidence rate of medical consultation for fecundity problems—analysis of a prevalent cohort using competing risks. *Hum. Reprod.* **28** 2872–2879.
- ERTEFAIE, A., ASGHARIAN, M. and STEPHENS, D. (2014). Propensity score estimation in the presence of length-biased sampling: A non-parametric adjustment approach. *Stat* **3** 83–94. [MR4027329](#) <https://doi.org/10.1002/sta.4.46>
- FARLEY ORDOVENSKY STANIEC, J. and WEBB, N. J. (2007). Utilization of infertility services: How much does money matter? *Health Serv. Res.* **42** 971–989.
- FINE, J. P. and GRAY, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *J. Amer. Statist. Assoc.* **94** 496–509. [MR1702320](#) <https://doi.org/10.2307/2670170>
- FINE, J. P., JIANG, H. and CHAPPELL, R. (2001). On semi-competing risks data. *Biometrika* **88** 907–919. [MR1872209](#) <https://doi.org/10.1093/biomet/88.4.907>
- CENTERS FOR DISEASE CONTROL (2014). National Public Health Action Plan for the Detection, Prevention, and Management of Infertility, Atlanta, Georgia: Centers for Disease Control and Prevention. Available at <http://www.cdc.gov/reproductivehealth/infertility/publichealth.htm>.
- HE, K., LI, Y., RAO, P. S., SUNG, R. S. and SCHAUBEL, D. E. (2020). Prognostic score matching methods for estimating the average effect of a non-reversible binary time-dependent treatment on the survival function. *Lifetime Data Anal.* **26** 451–470. [MR4102819](#) <https://doi.org/10.1007/s10985-019-09485-x>
- HERNÁN, M. A., COLE, S. R., MARGOLICK, J., COHEN, M. and ROBINS, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiol. Drug Saf.* **14** 477–491.
- KEIDING, N. (1991). Age-specific incidence and prevalence: A statistical perspective. *J. Roy. Statist. Soc. Ser. A* **154** 371–412. [MR1144166](#) <https://doi.org/10.2307/2983150>
- KEIDING, N., KVIST, K., HARTVIG, H., TVEDE, M. and JUUL, S. (2002). Estimating time to pregnancy from current durations in a cross-sectional sample. *Biostatistics* **3** 565–578.
- KEIDING, N., HANSEN, O. K. H., SØRENSEN, D. N. and SLAMA, R. (2012). The current duration approach to estimating time to pregnancy. *Scand. J. Stat.* **39** 185–204. [MR2927018](#) <https://doi.org/10.1111/j.1467-9469.2012.00798.x>
- LI, Y., SCHAUBEL, D. E. and HE, K. (2014). Matching methods for obtaining survival functions to estimate the effect of a time-dependent treatment. *Stat. Biosci.* **6** 105–126.
- MANDEL, M. (2010). The competing risks illness-death model under cross-sectional sampling. *Biostatistics* **11** 290–303. <https://doi.org/10.1093/biostatistics/kxp048>
- MANDEL, M. (2015). Analyzing multiple cross-sectional samples with application to hospitalization time after surgeries. *Stat. Med.* **34** 3415–3423. [MR3412641](#) <https://doi.org/10.1002/sim.6535>
- MANDEL, M. and FLUSS, R. (2009). Nonparametric estimation of the probability of illness in the illness-death model under cross-sectional sampling. *Biometrika* **96** 861–872. [MR2767275](#) <https://doi.org/10.1093/biomet/asp046>
- MANDEL, M. and RINOTT, Y. (2014). Estimation from cross-sectional samples under bias and dependence. *Biometrika* **101** 719–725. [MR3254912](#) <https://doi.org/10.1093/biomet/asu013>
- MCLAIN, A. C., SUNDARAM, R., THOMA, M. and BUCK LOUIS, G. M. (2014). Semiparametric modeling of grouped current duration data with preferential reporting. *Stat. Med.* **33** 3961–3972. [MR3261055](#) <https://doi.org/10.1002/sim.6216>
- MCLAIN, A. C., GUO, S., THOMA, M. and ZHANG, J. (2021). Supplement to “Length-biased semicompeting risks models for cross-sectional data: An application to current duration of pregnancy attempt data.” <https://doi.org/10.1214/20-AOAS1428SUPPA>, <https://doi.org/10.1214/20-AOAS1428SUPPB>
- R CORE TEAM (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- ROBINS, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association; Section on Bayesian Statistical Science* 6–10.
- SCHEIKE, T. H. and KEIDING, N. (2006). Design and analysis of time-to-pregnancy. *Stat. Methods Med. Res.* **15** 127–140. [MR2226588](#) <https://doi.org/10.1191/0962280206sm435oa>
- SIMONEAU, G., MOODIE, E. E. M., NIJJAR, J. S. and PLATT, R. W. (2020). Scottish early rheumatoid arthritis inception cohort investigators estimating optimal dynamic treatment regimes with survival outcomes. *J. Amer. Statist. Assoc.* **115** 1531–1539. [MR4143483](#) <https://doi.org/10.1080/01621459.2019.1629939>
- SLAMA, R., HANSEN, O. K., DUCOT, B., BOHET, A., SØRENSEN, D., GIORGIS, A. L., EIJKEMANS, M. J. C., ROSETTA, L., THALABARD, J. C. et al. (2012). Estimation of the frequency of involuntary infertility on a nation-wide basis. *Mol. Hum. Reprod.* **27** 1489–1498.
- SMARR, M. M., SAPRA, K. J., GEMMILL, A., KAHN, L. G., WISE, L. A., LYNCH, C. D., FACTOR-LITVAK, P., MUMFORD, S. L., SKAKKEBAEK, N. E. et al. (2017). Is human fecundity changing? A discussion of research and data gaps precluding us from having an answer. *Hum. Reprod.* **32** 499.

- TSIATIS, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proc. Natl. Acad. Sci. USA* **72** 20–22. [MR0356425](#) <https://doi.org/10.1073/pnas.72.1.20>
- VAN ES, B., KLAASSEN, C. A. J. and OUDSHOORN, K. (2000). Survival analysis under cross-sectional sampling: Length bias and multiplicative censoring. *J. Statist. Plann. Inference* **91** 295–312. [MR1814785](#) [https://doi.org/10.1016/S0378-3758\(00\)00183-X](https://doi.org/10.1016/S0378-3758(00)00183-X)
- VARDI, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika* **76** 751–761. [MR1041420](#) <https://doi.org/10.1093/biomet/76.4.751>
- WEINBERG, C. R. and GLADEN, B. C. (1986). The beta–geometric distribution applied to comparative fecundability studies. *Biometrics* **42** 547–560.
- YAMAGUCHI, K. (2003). Accelerated failure–time mover–stayer regression models for the analysis of last–episode data. *Sociol. Method.* **33** 81–110.

A BRIDGING MODEL TO RECONCILE STATISTICS BASED ON DATA FROM MULTIPLE SURVEYS

BY ANDREEA L. ERCIULESCU^{1,*}, JEAN D. OPSOMER^{1,†} AND F. JAY BREIDT²

¹Westat, * AndreeaErciulescu@westat.com; † Jean.Opsomer@westat.com

²Department of Statistics, Colorado State University, FJay.Breidt@colostate.edu

Surveys designed to collect data on similar variables using samples representing the same population may still result in different estimates due, for example, to differences in sample designs or modes of data collection. Considered in this paper is the case where two surveys were conducted concurrently, with one using the same methodology as used in prior rounds of the survey and the other using an updated methodology, resulting in substantial differences in several key estimates. Due to differences in sample size, only the latter survey was detailed enough for disaggregated-level estimates of publishable quality. We propose a hierarchical model to account for discrepancies in the estimates from the two surveys and a Bayesian approach for producing reliable estimates at various levels of aggregation. The model relies on a common latent structure at the disaggregated level to allow “bridging” between the two surveys. The methodology is applied to the 2016 National Survey of Fishing, Hunting and Wildlife-Associated Recreation and the 2016 50-State Surveys of Fishing, Hunting and Wildlife-Related Recreation. Aligning these two surveys is critical to extend the series of related statistics that have been published since 1955, allowing for meaningful comparisons over time despite the change in survey methodology.

REFERENCES

- ASH, S. (2014). Using successive difference replication for estimating variances. *Surv. Methodol.* **40** 47–60.
- BALGOBIN, N., BERG, E. and BARBOZA, W. (2014). A hierarchical Bayesian model for forecasting state-level corn yield. *Environ. Ecol. Stat.* **21** 507–530. [MR3248537](https://doi.org/10.1007/s10651-013-0266-z) <https://doi.org/10.1007/s10651-013-0266-z>
- ERCIULESCU, A. L., CRUZE, N. B. and NANDRAM, B. (2020). Statistical challenges in combining survey and auxiliary data to produce official statistics. *J. Off. Stat.* 63–88.
- ERCIULESCU, A. L., OPSOMER, J. D. and BREIDT, F. J. (2021). Supplement to “A bridging model to reconcile statistics based on data from multiple surveys.” <https://doi.org/10.1214/20-AOAS1437SUPP>.
- FAY, R. E. III and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277. [MR0548019](https://doi.org/10.1080/016214506000001293)
- GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–511.
- KOTT, P. (2001). The delete-a-group jackknife. *J. Off. Stat.* **17** 521–526.
- LOHR, S. L. and BRICK, J. M. (2012). Blending domain estimates from two victimization surveys with possible bias. *Canad. J. Statist.* **40** 679–696. [MR2998856](https://doi.org/10.1002/cjs.11153) <https://doi.org/10.1002/cjs.11153>
- OLSON, K., SMYTH, J. D., HORWITZ, R., KEETER, S., LESSER, V., MARKEN, S., MATHIOWETZ, N., MCCARTHY, J., O'BRIEN, E. et al. (2020). Transitions from telephone surveys to self-administered and mixed-mode surveys: AAPOR task force report. *Journal of Survey Statistics and Methodology* smz062.
- RAGHUNATHAN, T. E., XIE, D., SCHENKER, N., PARSONS, V. L., DAVIS, W. W., DODD, K. W. and FEUER, E. J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *J. Amer. Statist. Assoc.* **102** 474–486. [MR2370848](https://doi.org/10.1198/016214506000001293) <https://doi.org/10.1198/016214506000001293>
- ROCKVILLE INSTITUTE (2018). 2016 50-State Survey of Fishing, Hunting, and Wildlife-Related Recreation.
- STAN DEVELOPMENT TEAM (2020). Stan user’s guide.
- U. S. DEPARTMENT OF THE INTERIOR, FISH AND WILDLIFE SERVICE AND U. S. DEPARTMENT OF COMMERCE, CENSUS BUREAU (2018). 2016 National Survey of Fishing, Hunting, and Wildlife-Associated Recreation.

VAN DEN BRAKEL, J., ZHANG, X. and TAM, S.-M. (2020). Measuring discontinuities in time series obtained with repeated sample surveys. *Int. Stat. Rev.* **88** 155–175. [MR4088015](#) <https://doi.org/10.1111/insr.12347>

The Annals of Applied Statistics

Next Issues

- A multivariate spatio-temporal change point model of opioid overdose deaths in Ohio
STACI A. HEPLER, LANCE WALLER AND DAVID KLINE
- Unsupervised streaming anomaly detection for instrumented infrastructure
HENRIQUE HELFER HOELTGEBAUM, NIAL ADAMS AND DIN-HOUN LAU
- Function-on-function regression for the identification of epigenetic regions exhibiting windows of susceptibility to environmental exposures MICHELE ZEMPLÉNYI,
MARK J. MEYER, ANDRES CARDENAS, MARIE-FRANCE HIVERT,
SHERYL L. RIFAS-SHIMAN, HEIKE GIBSON, ITAI KLOOG, JOEL SCHWARTZ,
EMILY OKEN, DAWN L. DEMEO, DIANE R. GOLD AND BRENT A. COULL
- Two-stage circular-circular regression with zero-inflation: Application to medical sciences
JAYANT JHA AND PRAJAMITRA BHUYAN
- Joint and individual analysis of breast cancer histologic images and genomic covariates
IAIN CARMICHAEL, BENJAMIN C. CALHOUN, KATHERINE A. HOADLEY,
MELISSA A. TROESTER, JOSEPH GERADTS, HEATHER D. COUTURE, LINNEA OLSSON,
CHARLES M. PEROU, MARC NIETHAMMER, JAN HANNIG AND J. S. MARRON
- Perturbed factor analysis: Accounting for group differences in exposure profiles
ARKAPRAVA ROY, ISAAC LAVINE, AMY H. HERRING AND DAVID B. DUNSON
- Global estimation and scenario-based projections of sex ratio at birth and missing female births using a Bayesian hierarchical time series mixture model FENGQING CHAO,
PATRICK GERLAND, ALEX R. COOK AND LEONTINE ALKEMA
- Targeted Smooth Bayesian Causal Forests: An analysis of heterogeneous treatment effects for simultaneous versus interval medical abortion regimens over gestation
JENNIFER E. STARLING, JARED S. MURRAY, PATRICIA A. LOHR,
ABIGAIL R. A. AIKEN, CARLOS M. CARVALHO AND JAMES G. SCOTT
- Partial-mastery cognitive diagnosis models ZHUORAN SHANG,
ELENA A. EROSHEVA AND GONGJUN XU
- A Bayesian nonparametric approach to super-resolution single-molecule localization
MARIANO IGNACIO GABITTO, HERVE MARIE-NELLY, ARI PAKMAN,
ANDRAS PATAKI, XAVIER DARZACQ AND MICHAEL JORDAN
- Zero-inflated quantile rank-score based test (ZIQRank) with application to scRNA-seq differential gene expression analysis WODAN LING, WENFEI ZHANG,
BIN CHENG AND YING WEI
- Sparse matrix linear models for structured high-throughput data
JANE W. LIANG AND SAUNAK SEN
- Modeling past event feedback through biomarker dynamics in the multi-state event analysis for cardiovascular disease data CHUOXIN MA, HONGSHENG DAI AND JIANXIN PAN
- Estrogen receptor expression on breast cancer patients' survival under shape restricted Cox regression model JING QIN, GENG DENG, JING NING, AO YUAN, AND YU SHEN
- Markov random field models for vector-based representations of landscapes
PATRIZIA ZAMBERLETTI, JULIEN PAPAÏX, EDITH GABRIEL AND THOMAS OPITZ
- The information in covariate imbalance in studies of hormone replacement therapy
DYLAN SMALL, RUOQI YU AND PAUL ROSENBAUM

Continued

The Annals of Applied Statistics

Next Issues—Continued

- Information content of high-order associations of the human gut microbiota network
WESTON D. VILES, JULIETTE C. MADAN, HONGZHE LI,
MARGARET R. KARAGAS AND ANNE G. HOEN
- Analysing the causal effect of London cycle superhighways on traffic congestion
PRAJAMITRA BHUYAN, EMMA J. MCCOY, HAOJIE LI AND DANIEL J. GRAHAM
- Assessing selection bias in regression coefficients estimated from non-probability samples, with
applications to genetics and demographic surveys BRADY THOMAS WEST,
RODERICK J. A. LITTLE, REBECCA R. ANDRIDGE, PHILIP S. BOONSTRA,
ERIN B. WARE, ANITA PANDIT AND FERNANDA ALVARADO-LEITON
- Spatial voting models in circular spaces XINGCHEN YU AND ABEL RODRÍGUEZ
- Identifying the recurrence of sleep apnea using a harmonic hidden Markov model
BENIAMINO HADI-AMAR, BÄRBEL FINKENSTÄDT,
MARK FIECAS AND ROBERT HUCKSTEPP
- Bayesian multi-study factor analysis for high-throughput biological data
ROBERTA DE VITO, RUGGERO BELLIO, LORENZO TRIPPA AND GIOVANNI PARMIGIANI
- Markov-switching state space models for uncovering musical interpretation
DANIEL J. McDONALD, MICHAEL MCBRIDE,
YUPENG GU AND CHRISTOPHER RAPHAEL
- RADIOHEAD: Radiogenomic analysis incorporating tumor heterogeneity in imaging through
densities SHARIQ MOHAMMED, KARTHIK BHARATH, SEBASTIAN KURTEK,
ARVIND RAO AND VEERABHADRAN BALADANDAYUTHAPANI
- Clustering on the torus by conformal prediction
SUNGKYU JUNG, KIHOO PARK AND BYUNGWON KIM
- Assessing the reliability of wind power operations under a changing climate with a
non-Gaussian bias correction JIACHEN ZHANG, PAOLA CRIPPA,
MARC GENTON AND STEFANO CASTRUCCIO
- Bayesian joint modeling of chemical structure and dose response curves
KELLY R. MORAN, DAVID DUNSON, MATTHEW W. WHEELER AND AMY H. HERRING
- Orthogonal subsampling for big data linear regression. LIN WANG
- Modelled approximations to the ideal filter with application to GDP and its components
THOMAS M. TRIMBUR AND TUCKER MCELROY
- Predicting competitions by combining conditional logistic regression and subjective Bayes:
An Academy Awards case study CHRISTOPHER T. FRANCK AND
CHRISTOPHER E. WILSON
- Qini-based uplift regression MOULOUD BELBAHRI, ALEJANDRO MURUA,
OLIVIER GANDOUET AND VAHID PARTOVI NIA
- Simultaneous non-Gaussian component analysis (SING) for data integration in neuroimaging
BENJAMIN B. RISK AND IRINA GAYNANOVA
- Mediation analysis for associations of categorical variables: The role of education in social class
mobility in Britain JOUNI KUHA, ERZSEBET BUKODI AND JOHN H. GOLDTHORPE
- Subgroup identification and variable selection for treatment decision making
BAQUN ZHANG AND MIN ZHANG

Continued

The Annals of Applied Statistics

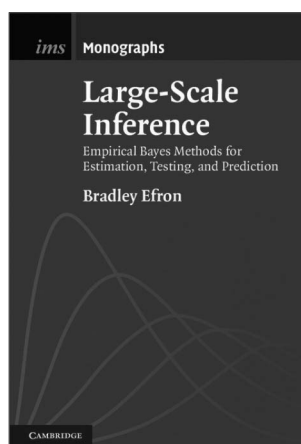
Next Issues—Continued

- Bridging randomized controlled trials and single-arm trials using commensurate priors in arm-based network meta-analysis . . . ZHENXUN WANG, LIFENG LIN, THOMAS MURRAY, JAMES S. HODGES AND HAITAO CHU
- Pan-disease clustering analysis of the trend of period prevalence . . . SNEHA JADHAV, CHENJIN MA, YEFEI JIANG, BENCHANG SHIA AND SHUANGGE MA
- Improving exoplanet detection power: Multivariate Gaussian process models for stellar activity . . . DAVID EDWARD JONES, DAVID C. STENNING, ERIC B. FORD, ROBERT L. WOLPERT, THOMAS J. LOREDO, CHRISTIAN GILBERTSON AND XAVIER DUMUSQUE
- Estimating animal utilization distributions from multiple data types: A joint spatio-temporal point process framework . . . JOE WATSON, RUTH JOY, DOMINIC TOLLIT, SHEILA J. THORNTON AND MARIE AUGER-MÉTHÉ
- Diagnosis-group-specific transitional care program recommendations for thirty-day rehospitalization reduction . . . MENGANG YU, CHENSHENG KUANG, JARED DAVIS HULING AND MAUREEN A. SMITH
- Space-time smoothing models for sub-national measles routine immunization coverage estimation with complex survey data . . . TRACY QI DONG AND JON WAKEFIELD
- Tensor quantile regression with application to association between neuroimages and human intelligence . . . CAI LI AND HEPING ZHANG
- Bounding the local average treatment effect in an instrumental variable analysis of engagement with a mobile intervention . . . ANDREW JUSTIN SPIEKER, ROBERT GREEVY, LYNDSEY NELSON AND LINDSAY MAYBERRY
- A functional-data approach to the Argo data . . . DREW YARGER, STILIAN STOEV AND TAIEN HSING
- Inferring food intake from multiple biomarkers using a latent variable model . . . SILVIA D'ANGELO, LORRAINE BRENNAN AND ISOBEL CLAIRE GORMLEY



The Institute of Mathematical Statistics presents

IMS MONOGRAPHS



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

Bradley Efron

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

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

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

www.cambridge.com/ims

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