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
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 GIACOFICI 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 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.

Copyright © 2021 by the Institute of Mathematical Statistics
Printed in the United States of America
A HERMITE–GAUSSIAN BASED EXOPLANET RADIAL VELOCITY ESTIMATION METHOD

BY PARKER H. HOLZER¹, JESSI CISIEWSKI-KHE², DEBRA FISCHER³,* AND LILY ZHAO³,†

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 rrvmethod, 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., ADRIANZKY, G., KNISPEL, 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.
BEDELL, 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.

Key words and phrases. Exoplanets, Hermite–Gaussian functions, linear regression, astrostatistics, Doppler effect.
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)

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.

Ciuryla, 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 Dyn.* **84** 1157–1161. [MR3486561](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., Segransan, 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 plages. *Astrophys. J.* **796** 132.

Dumusque, X., Borsa, F., Damasso, M., Diaz, 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., Arriaga, P., Balve, 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., Terrrien, R., Mahadevan, S., Roy, A., Bender, C., Stefansson, 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-AOAS1406SUPP](https://doi.org/10.1214/20-AOAS1406SUPP)

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](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)

Jurgenson, C., Fischer, D., McCracken, T., Sawyer, D., Szymkowiak, 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-1. 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 JERINGAN, 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. *Impact of plages. 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 coraille 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.

RAJPAL, 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.

RAJPAL, 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.

RIFTEL, R. A. (2010). Profit: A new alternative for emission-line pro file fit ting. *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 CISIEWSKIKEHE, 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 Sobek, 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., CISIEWSKIKEHE, 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. KEELER2,∗ AND SUMMER RUPPER2,†

1Department of Statistics, Brigham Young University, pwhite@stat.byu.edu
2Department 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., FAHNSTOCK, M. and SCAMBOS, T. (2004). Extreme firm metamorphism: Impact of decades of vapor transport on near-surface firn 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.

Key words and phrases. Bayesian statistics, Gaussian process, monotonic regression, spatial statistics, spline.
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 firm 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†

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.

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

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

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

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

CASTRO-CAMILLO, 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

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.
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. Guttorf 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 Travelletti2, *, Jo Eidsvik3, David Ginsbourger2, † and Kanna Rajan4

1 Department of Marine Technology, Norwegian University of Science and Technology (NTNU), trygve.o.fossum@ntnu.no
2 Institute of Mathematical Statistics and Actuarial Science, University of Bern, *cedric.travelletti@stat.unibe.ch; †david.ginsbourger@stat.unibe.ch
3 Department of Mathematical Sciences, NTNU, jo.eidsvik@ntnu.no
4 Underwater 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 RICHE, 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., PICHEY, V. and VÁZQUEZ, 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

BELLENGHAM, J. G. and RAJAN, K. (2007). Robotics in remote and hostile environments. Science 318 1098–1102.

BHATTACHARJYA, 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.
WACKERNAGEL, H. (2003). Multivariate Geostatistics: An Introduction with Applications. Springer, Berlin.

WIKLE, C. K., MILLIFF, R. F., HERBEI, 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

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

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.

Dreyden, 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

Dreyden, 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

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

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

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

Joshi, S. H., Klassen, E., Srivastava, A. and Jermy, I. H. (2007). A novel representation for Riemannian analysis of elastic curves in $\mathbb{R}^p$. In IEEE Conference on Computer Vision and Pattern Recognition 1–7.

Key words and phrases. Dimension reduction, LDA, pairwise classification, projection point, QDA, registration.
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., Panjivay, 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

Vailant, 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 to 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.
SIMULTANEOUS INFERENCE OF PERIODS AND PERIOD-LUMINOSITY RELATIONS FOR MIRA VARIABLE STARS

BY SHIYUAN HE, ZHENFENG LIN, WENLONG YUAN, LUCAS M. MACRI AND JIANHUA Z. HUANG

1Institute of Statistics and Big Data, Renmin University of China, heshiyuan@ruc.edu.cn
2Microsoft, zhenfeng.lin@microsoft.com
3Department of Physics and Astronomy, Johns Hopkins University, wyuan10@jhu.edu
4Department of Physics and Astronomy, Texas A&M University, lmacri@tamu.edu
5Department 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

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

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

GIEREN, W., GORSKI, M., PIETRZYNSKI, 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.

Key words and phrases. Variational inference, hierarchical Bayesian modeling, astrostatistics, Mira variables.
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)

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., NSENGYIUMVA, F. and MATSUNAGA, N. (2014). Vizier online data catalog: Jhks photometry of agb stars in ngc 6822 (whiteclock+, 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

¹Division of the National Toxicology Program, National Institute of Environmental Health Sciences, kyle.messier@nih.gov
²Department 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
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 PM2.5 in the contiguous United States. Environ. Sci. Technol. 47 7233–7241. https://doi.org/10.1021/es400039u.
BOERSMA, K. F., ESKE, H. J., VEERKIND, 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

Key words and phrases. General Vecchia approximation, spatial statistics, Gaussian process, variable selection, air pollution, Kriging.
WEISS, D. J., NELSON, A., GIBSON, H. S., TEMPERLEY, W., PEEDELL, S., LIEBER, A., HANCHER, M., POUART, 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

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

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

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
PROBABILISTIC FORECASTING OF THE ARCTIC SEA ICE EDGE WITH CONTOUR MODELING

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

¹Department of Statistics, University of Washington, direch@uw.edu
²Departments of Statistics and Sociology, University of Washington, raftery@uw.edu
³Department 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
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
DIRECTOR, H. M., RAFTERY, A. E. and BITZ, C. M. (2017). Improved sea ice forecasting through spatiotemporal bias correction. J. Climate 30 9493–9510.

Key words and phrases. Spatiotemporal, climate change, forecasting, postprocessing, mixtures.
GUEMAS, V., BLANCHARD-WRIGGLESWORTH, E., CHEVALLIER, M., DAY, J. J., DEQUÉ, M., DOBLAS-REYES, F. J., FÜCKAR, 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., BALMASEDA, M. A., MOLTENI, F., MAGNUSSON, L., TIETSCHIE, 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.

TIETSCHIE, 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.

WAYNAND, 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 Artic 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 regridder 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 CAPEZZA1,*, BIAGIO PALUMBO1,†, YANNIG GOUDE2, SIMON N. WOOD3 AND MATTEO FA SI OLO4

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 FASI OLO, 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 FASI OLO, 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.
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 ALTSMANN, 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/ars013

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

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.
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., RICHÉ, R. L. and CARRARO, L. (2008). A multi-points criterion for deterministic parallel global optimization based on Gaussian processes. HAL Preprint Hal-00260579.

Key words and phrases. Design of metamaterials, multitarget design, response surface learning, design of computer experiments, Bayesian optimization.
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

BENGTSSON, H. (2017). doFuture: A universal foreach parallel adaptor using the future API of the ‘future’ package. R package version 0.6.0.

BENGSTSSON, 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

Key words and phrases. Animal telemetry, semi-Markov model, movement model, multiple imputation, northern fur seal.
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 AVRIL³,* AND JONAS WALDENSTRÖM³,†

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. (2014). Bayesian estimation of abundance for open populations with covariate dependent capture and survival probabilities draft. Available at http://www.simon.bonners.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

BROOKS, 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 WinBUGS. *J. Stat. Softw.* 14 1–24.
CRAPE, 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.

FILIP, 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

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

GUNNARSSON, G., LATORRE-MARGALEF, N., HOBBIN, 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

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

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

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

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

SCHAUB, M., JENN, 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., WIEBEL, 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\(^1,\ast\), ZHAO REN\(^1,\dagger\), JUAN C. CEBEDÓN\(^2,\ddagger\) AND WEI CHEN\(^2,\S\)

\(^1\)Department of Statistics, University of Pittsburgh, \(*\) roz16@pitt.edu; \(^\dagger\) zren@pitt.edu

\(^2\)Department of Pediatrics, UPMC Children’s Hospital of Pittsburgh, University of Pittsburgh, \(^\ddagger\) juan.celedon@chp.edu; \(^\S\) 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

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

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

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. J. Roy. Statist. Soc. Ser. B 36 192–236. MR0373208

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

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.
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., CLEEDON, 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. BARTLETT1,*, IOANNIS KOSMIDIS2,3 AND RICARDO SILVA1,3

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., MA-rINE, 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

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

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

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

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

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

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9 432–441.

Key words and phrases. Bayesian inference, sparse statistical models, time-varying networks, genomic networks.
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

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., LI-VAK, 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 MA1,*, KEEGAN KORTHAUER2, CHRISTINA KENDZIORSKI3 AND MICHAEL A. NEWTON1,†

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., TIECHMANN, 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., KOWASKY, 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.
LARGE-SCALE MULTIPLE INFERENCE OF COLLECTIVE DEPENDENCE WITH APPLICATIONS TO PROTEIN FUNCTION

BY ROBERT JERNIGAN¹,*, KEJUE JIA¹,†, 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 Large inference procedure to DETect significant \( k \)-COmBorative DEpendence 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

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

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

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.jmeth.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

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

Key words and phrases. Collective dependence, false discovery rate, information theoretic measure, multiple testing, protein coevolution, structural biology.
A BAYESIAN NONPARAMETRIC MODEL FOR INFERRING SUBCLONAL POPULATIONS FROM STRUCTURED DNA SEQUENCING DATA

BY SHAI HE**1,*, AARON SCHEIN**2, VISHAL SARSANI1,† AND PATRICK FLAHERTY1,‡

1Department of Mathematics and Statistics, University of Massachusetts Amherst, *shaihe@math.umass.edu; †vsarsani@umass.edu; ‡pflaherty@umass.edu
2Data 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

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., CAVENE, 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., TREV, D., ENDRIS, V., ISMAEL, F., BANGEMANN, N., BLOMER, 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.
TREUTLEIN, B., BROWNFIELD, D. G., WU, A. R., NEFF, N. F., MANTALAS, G. L., ESPINOZA, F. H., DESAI, 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
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

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

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

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

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

Friedman, J., Hastie, T. and Tibshirani, R. (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

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

Key words and phrases. Mixture regression model, prediction, variable selection, heterogeneous data, spectrometry data.
EFFICIENT BAYESIAN INFERENCE OF GENERAL GAUSSIAN MODELS ON LARGE PHYLOGENETIC TREES

BY PAUL BASTIDE¹, LAM SI TUNG HO², GUY BAELE³,*, PHILIPPE LEMEY³,† 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
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
ALIZON, S., VON WYL, V., STADLER, T., KOUYOS, R. D., YERLY, S., HIRSCHEL, 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.
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., Lemy, 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 Lynch, 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

Lemy, 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., Alfaran, A. H., Alquraish, 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
Wiens, J. J., Kuczyński, 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). `<scp>ggtree</scp>`: An `<scp>r</scp>` 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 SUN²,* AND LI HSU²,†

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.

BUINIELLO, 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.

BUINIELLO, 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

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.
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 SHABBIR, 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

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

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 TOUKOUSHIAN, 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

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

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

COHN, E. (1973). Factors affecting variations in faculty salaries and compensation in institutions of higher education. The Journal of Higher Education 44 124–136.

DELOMME, 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

Key words and phrases. Faculty salary data, mixture models, cluster analysis, tensor data.
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 McLAUGHLIN, 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

VIROLI, C. (2011b). Model based clustering for three-way data structures. *Bayesian Anal.* 6:573–602. MR2869958 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

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

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

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

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

Key words and phrases. Homelessness, housing affordability, Bayesian hierarchical model, random partition model, Ewens–Pitman attraction.
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. MCLAIN1,*, SIYUAN GUO1,†, MARIE THOMA2,§ AND JIAJIA ZHANG1,‡

1Department of Epidemiology and Biostatistics, University of South Carolina, *mclaina@mailbox.sc.edu; †siyuang@email.sc.edu; ‡jzhang@mailbox.sc.edu
2Department 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

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/10.1198/01621450800000382

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

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

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

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

Key words and phrases. Cross-sectional data, infertility, length-biased survival, semicompeting risks.
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/sta4.46

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 SCHAUDEL, 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., SCHAUDEL, 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

MCLAIRN, 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

MCLAIRN, 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 inference 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., EHKEIMANS, 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

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. ERCIULESCU1,*, JEAN D. OPSOMER1,† AND F. JAY BREIDT2

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
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
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
OLSON, K., SMYTH, J. D., HORWITZ, R., KEESTER, 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
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.

Key words and phrases. Combining data, disaggregation, fishing participation, hierarchical Bayes.
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
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, Niall Adams and Din-Houn Lau

Function-on-function regression for the identification of epigenetic regions exhibiting windows of susceptibility to environmental exposures. Michele Zemplenyi, Mark J. Meyer, Andres Cardenas, Marie-France Hivert, Sheryl L. Rifas-Shiman, Heike Gibson, Itai Kloo, 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 Prajimatra 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 Olson, Charles M. Perou, Marc Niethammer, Jan Hannig and J. S. Marron

Perturbed factor analysis: Accounting for group differences in exposure profiles
Arkapra 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. 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, András 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 Papaix, 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 HADJ-AMAR, BÄRBEL FINKENSTÄDT, MARK PIECAS AND ROBERT HUCKSTEDD

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, KIHO 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

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 GAYNANAVA

Mediation analysis for associations of categorical variables: The role of education in social class mobility in Britain
JOUNI KUHA, ERZSEBET BUKODI AND JOHN H. GOLDSBOROUGH

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, YEEFEI 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 ............... MENGGANG 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, LYNDsay NELSON AND LINDSAY MAYBERRY

A functional-data approach to the Argo data

DREW YARGER, STILIAN STOEV AND TAILEN HSING

Inferring food intake from multiple biomarkers using a latent variable model

SILVIA D’ANGELO, LORRAINE BRENNAN AND ISOBEL CLAIRE GORMLEY
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.