ISSN 1932-6157 (print)
ISSN 1941-7330 (online)

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles

continued

ISSN 1932-6157 (print)
ISSN 1941-7330 (online)

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles—*Continued from front cover*

Vol. 11, No. 3—September 2017

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1193–1216
DOI: 10.1214/17-AOAS1058

THE PROBLEM OF INFRA-MARGINALITY IN OUTCOME TESTS FOR DISCRIMINATION

BY CAMELIA SIMOIU, SAM CORBETT-DAVIES AND SHARAD GOEL

Stanford University

Outcome tests are a popular method for detecting bias in lending, hiring, and policing decisions. These tests operate by comparing the success rate of decisions across groups. For example, if loans made to minority applicants are observed to be repaid more often than loans made to whites, it suggests that only exceptionally qualified minorities are granted loans, indicating discrimination. Outcome tests, however, are known to suffer from the problem of *infra-marginality*: even absent discrimination, the repayment rates for minority and white loan recipients might differ if the two groups have different risk distributions. Thus, at least in theory, outcome tests can fail to accurately detect discrimination. We develop a new statistical test of discrimination—the threshold test—that mitigates the problem of infra-marginality by jointly estimating decision thresholds and risk distributions. Applying our test to a dataset of 4.5 million police stops in North Carolina, we find that the problem of infra-marginality is more than a theoretical possibility, and can cause the outcome test to yield misleading results in practice.

REFERENCES

ALPERT, G. P., SMITH, M. R. and DUNHAM, R. G. (2004). Toward a better benchmark: Assessing the utility of not-at-fault traffic crash data in racial profiling research. *Justice Res. Policy* **6** 43–69.

ANTONOVICS, K. and KNIGHT, B. G. (2009). A new look at racial profiling: Evidence from the Boston police department. *Rev. Econ. Stat.* **91** 163–177.

ANWAR, S. and FANG, H. (2006). An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *Am. Econ. Rev.* **96** 127–151.

ARROW, K. (1973). The theory of discrimination. In *Discrimination in Labor Markets* Princeton Univ. Press, Princeton.

AYRES, I. (2002). Outcome tests of racial disparities in police practices. *Justice Res. Policy* **4** 131–142.

BECKER, G. S. (1957). *The Economics of Discrimination*. Univ. Chicago Press, Chicago, IL.

BECKER, G. S. (1993). Nobel lecture: The economic way of looking at behavior. *J. Polit. Econ.* **101** 385–409.

CARPENTER, B., GELMAN, A., HOFFMAN, M., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M. A., GUO, J., LI, P. and STAN, A. R. (2016). A probabilistic programming language. *J. Stat. Softw.*

CARR, J. H. and MEGBOLUGBE, I. F. (1993). The Federal Reserve Bank of Boston study on mortgage lending revisited. *J. Hous. Res.* **4** 277–313.

CORBETT-DAVIES, S., PIERSON, E., FELLER, A., GOEL, S. and HUQ, A. (2017). Algorithmic decision making and the cost of fairness. Preprint. Available at 1701.08230.

Key words and phrases. Tests for discrimination, outcome test, benchmark test, infra-marginality, traffic stops, policing.

DUANE, S., KENNEDY, A. D., PENDLETON, B. J. and ROWETH, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195** 216–222.

ENGEL, R. S. and CALNON, J. M. (2004). Comparing benchmark methodologies for police-citizen contacts: Traffic stop data collection for the Pennsylvania State Police. *Police Q.* **7** 97–125.

ENGEL, R. S. and TILLYER, R. (2008). Searching for equilibrium: The tenuous nature of the outcome test. *Justice Q.* **25** 54–71.

EPP, C. R., MAYNARD-MOODY, S. and HAIDER-MARKEL, D. P. (2014). *Pulled over*: *How Police Stops Define Race and Citizenship*. Univ. Chicago Press, Chicago, IL.

GALSTER, G. C. (1993). The facts of lending discrimination cannot be argued away by examining default rates. *Hous. Policy Debate* **4** 141–146.

GELMAN, A., FAGAN, J. and KISS, A. (2007). An analysis of the New York City Police Department's "stop-and-frisk" policy in the context of claims of racial bias. *J. Amer. Statist. Assoc.* **102** 813–823. MR2411646

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. MR1422404

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2027492

GOEL, S., RAO, J. M. and SHROFF, R. (2016). Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *Ann. Appl. Stat.* **10** 365–394. MR3480500

GOEL, S., PERELMAN, M., SHROFF, R. and SKLANSKY, D. (2017). Combatting police discrimination in the age of big data. *New Crim. Law Rev.* **20** 181–232.

GROGGER, J. and RIDGEWAY, G. (2006). Testing for racial profiling in traffic stops from behind a veil of darkness. *J. Amer. Statist. Assoc.* **101** 878–887. MR2324089

HETEY, R., MONIN, B., MAITREYI, A. and EBERHARDT, J. (2016). Data for change: A statistical analysis of police stops, searches, handcuffings, and arrests in oakland, Calif., 2013-2014. Technical report, Stanford University, SPARQ: Social Psychological Answers to Real-World Questions.

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

JORDAN, M. I. (2004). Graphical models. *Statist. Sci.* **19** 140–155. MR2082153

KNOWLES, J., PERSICO, N. and TODD, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *J. Polit. Econ.* **109** 203–229.

LANGE, J. E., BLACKMAN, K. O. and JOHNSON, M. B. (2001). Speed violation survey of the New Jersey turnpike: Final report, Public Services Research Institute.

MACLIN, T. (2008). Good and bad news about consent searches in the Supreme Court. *McGeorge Law Rev.* **39** 27.

MCCONNELL, E. H. and SCHEIDEGGER, A. R. (2001). Race and speeding citations: Comparing speeding citations issued by air traffic officers with those issued by ground traffic officers. In *Annual Meeting of the Academy of Criminal Justice Sciences*, Washington, DC.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

NEAL, R. M. (1994). An improved acceptance procedure for the hybrid Monte Carlo algorithm. *J. Comput. Phys.* **111** 194–203. MR1271540

PHELPS, E. S. (1972). The statistical theory of racism and sexism. *Am. Econ. Rev.* **62** 659–661.

PIERSON, E., CORBETT-DAVIES, S. and GOEL, S. (2017). Fast threshold tests for detecting discrimination. Preprint. Available at 1702.08536.

RIDGEWAY, G. (2006). Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *J. Quant. Criminol.* **22** 1–29.

RIDGEWAY, G. and MACDONALD, J. M. (2009). Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *J. Amer. Statist. Assoc.* **104** 661–668. MR2751446

WALKER, S. (2003). Internal benchmarking for traffic stop data: An early intervention system approach. Technical report, Police Executive Research Forum.

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1217–1244
DOI: 10.1214/16-AOAS955

LATENT SPACE MODELS FOR MULTIVIEW NETWORK DATA

BY MICHAEL SALTER-TOWNSHEND[1] AND TYLER H. MCCORMICK[2]

University of Oxford and University of Washington

Social relationships consist of interactions along multiple dimensions. In social networks, this means that individuals form multiple types of relationships with the same person (e.g., an individual will not trust all of his/her acquaintances). Statistical models for these data require understanding two related types of dependence structure: (i) structure within each relationship type, or network view, and (ii) the association between views. In this paper, we propose a statistical framework that parsimoniously represents dependence between relationship types while also maintaining enough flexibility to allow individuals to serve different roles in different relationship types. Our approach builds on work on latent space models for networks [see, e.g., *J. Amer. Statist. Assoc.* **97** (2002) 1090–1098]. These models represent the propensity for two individuals to form edges as conditionally independent given the distance between the individuals in an unobserved social space. Our work departs from previous work in this area by representing dependence structure between network views through a multivariate Bernoulli likelihood, providing a representation of between-view association. This approach infers correlations between views not explained by the latent space model. Using our method, we explore 6 multiview network structures across 75 villages in rural southern Karnataka, India [Banerjee et al. (2013)].

REFERENCES

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E., XING, E. P. and JAAKKOLA, T. (2006). Mixed membership stochastic block models for relational data with application to protein-protein interactions. In *Proceedings of the International Biometrics Society Annual Meeting* 15.

AIROLDI, E., BLEI, D., FIENBERG, S. and XING, E. (2008). Mixed-membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

BANERJEE, A., CHANDRASEKHAR, A., DUFLO, E. and JACKSON, M. O. (2013). The Diffusion of Microfinance. The Abdul Latif Jameel Poverty Action Lab Dataverse.

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, Cambridge, MA. MR0381130

BUTTS, C. T. (2008). Social network analysis: A methodological introduction. *Asian J. Soc. Psychol.* **11** 13–41.

BUTTS, C. T. (2010). Sna: Tools for Social Network Analysis. Univ. California, Irvine. R package Version 2.1-0.

BUTTS, C. T. and CARLEY, K. M. (2005). Some simple algorithms for structural comparison. *Comput. Math. Organ. Theory* **11** 291–305.

CHANG, J. and BLEI, D. M. (2010). Hierarchical relational models for document networks. *Ann. Appl. Stat.* **4** 124–150. MR2758167

Key words and phrases. Latent space model, multiview relational data, social network.

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

DAI, B., DING, S. and WAHBA, G. (2013). Multivariate Bernoulli distribution. *Bernoulli* **19** 1465–1483. MR3102559

DIPRETE, T. A., GELMAN, A., MCCORMICK, T., TEITLER, J. and ZHENG, T. (2011). Segregation in social networks based on acquaintanceship and trust1. *Am. J. Sociol.* **116** 1234–1283.

FIENBERG, S. E., MEYER, M. M. and WASSERMAN, S. S. (1985). Statistical analysis of multiple sociometric relations. *J. Amer. Statist. Assoc.* **80** 51–67.

GOLLINI, I. and MURPHY, T. B. (2016). Joint modeling of multiple network views. *J. Comput. Graph. Statist.* **25** 246–265. MR3474046

GREENE, D. and CUNNINGHAM, P. (2013). Producing a unified graph representation from multiple social network views. In *Proceedings of the 5th Annual ACM Web Science Conference* (*WebSci'*13) 118–121.

HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354. MR2364300

HOFF, P. D. (2011a). Hierarchical multilinear models for multiway data. *Comput. Statist. Data Anal.* **55** 530–543. MR2736574

HOFF, P. D. (2011b). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Anal.* **6** 179–196. MR2806238

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262

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

HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76** 33–65. MR0608176

JACKSON, M. O., RODRIGUEZ-BARRAQUER, T. and TAN, X. (2012). Social capital and social quilts: Network patterns of favor exchange. *Am. Econ. Rev.* **102** 1857–1897.

KRIVITSKY, P. N., HANDCOCK, M. S., RAFTERY, A. E. and HOFF, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Soc. Networks* **31** 204–213.

MCCORMICK, T. H. and ZHENG, T. (2012). Latent demographic profile estimation in hard-to-reach groups. *Ann. Appl. Stat.* **6** 1795–1813. MR3058684

PATTISON, P. and WASSERMAN, S. (1999). Logit models and logistic regressions for social networks: II. Multivariate relations. *Br. J. Math. Stat. Psychol.* **52** 169–193.

ROBINS, G., PATTISON, P., KALISH, Y. and LUSHER, D. (2007). An introduction to exponential random graph (p^*) models for social networks. *Soc. Networks* **29** 173–191.

SALTER-TOWNSHEND, M. and MCCORMICK, T. H. (2017). Supplement to "Latent space models for multiview network data." DOI:10.1214/16-AOAS955SUPP.

SALTER-TOWNSHEND, M. and MURPHY, T. B. (2013). Variational Bayesian inference for the latent position cluster model for network data. *Comput. Statist. Data Anal.* **57** 661–671. MR2981116

SALTER-TOWNSHEND, M., WHITE, A., GOLLINI, I. and MURPHY, T. B. (2012). Review of statistical network analysis: Models, algorithms, and software. *Stat. Anal. Data Min.* **5** 260–264. MR2958152

SAMPSON, S. F. (1969). Crisis in a cloister. Unpublished doctoral dissertation, Cornell University.

STRAUSS, D. and IKEDA, M. (1990). Pseudolikelihood estimation for social networks. *J. Amer. Statist. Assoc.* **85** 204–212. MR1137368

STAN DEVELOPMENT TEAM (2013). Stan: A C++ library for probability and sampling. Version 2.

WAKEFIELD, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer, New York. MR3025225

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1245–1274
DOI: 10.1214/16-AOAS1014

A BAYESIAN APPROACH TO THE GLOBAL ESTIMATION OF MATERNAL MORTALITY[1]

BY LEONTINE ALKEMA[*], SANQIAN ZHANG[†], DORIS CHOU[‡], ALISON GEMMILL[§], ANN-BETH MOLLER[‡], DORIS MA FAT[‡], LALE SAY[‡], COLIN MATHERS[‡] AND DANIEL HOGAN[‡]

University of Massachusetts, Amherst[*], Harvard University[†], World Health Organization[‡] and University of California, Berkeley[§]

The maternal mortality ratio (MMR) is defined as the number of maternal deaths in a population per 100,000 live births. Country-specific MMR estimates are published on a regular basis by the United Nations Maternal Mortality Estimation Inter-agency Group (UN MMEIG) to track progress in reducing maternal deaths and were used to evaluate regional and national performance related to Millennium Development Goal (MDG) 5, which called for a 75% reduction in the MMR between 1990 and 2015.

Until 2014, the UN MMEIG used a multilevel regression model for producing estimates for countries without sufficient data from vital registration systems. While this model worked well in the past to assess MMR levels for countries with limited data, it was deemed unsatisfactory for final MDG 5 reporting for countries where longer time series of observations had become available because, by construction, estimated trends in the MMR were covariate-driven only and did not necessarily track data-driven trends.

We developed a Bayesian maternal mortality estimation model, which extends upon the UN MMEIG multilevel regression model. The new model assesses data-driven trends through the inclusion of an ARIMA time series model that captures accelerations and decelerations in the rate of change in the MMR. Varying reporting and data quality issues are accounted for in source-specific data models. The revised model provides data-driven estimates of MMR levels and trends and was used for MDG 5 reporting for all countries.

REFERENCES

ALKEMA, L. and NEW, J. R. (2014). Global estimation of child mortality using a Bayesian B-spline Bias-reduction model. *Ann. Appl. Stat.* **8** 2122–2149. MR3292491

ALKEMA, L., CHOU, D., HOGAN, D., ZHANG, S., MOLLER, A.-B., GEMMILL, A., FAT, D. M., BOERMA, T., TEMMERMAN, M., MATHERS, C. and SAY, L. (2016). Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: A systematic analysis by the un maternal mortality estimation inter-agency group. *The Lancet* **387** 462–474.

Key words and phrases. ARIMA time series models, Bayesian inference, multilevel regression model, maternal mortality ratio, Millennium Development Goal 5, UN Maternal Mortality Estimation Inter-agency Group (UN MMEIG).

ALKEMA, L., ZHANG, S., CHOU, D., GEMMILL, A., MOLLER, A.-B., FAT, D. M., SAY, L., MATHERS, C. and HOGAN, D. (2017). Supplement to "A Bayesian approach to the global estimation of maternal mortality." DOI:10.1214/16-AOAS1014SUPP.

CHAO, F. and ALKEMA, L. (2014). How informative are vital registration data for estimating maternal mortality? A Bayesian analysis of WHO adjustment data and parameters. *Statistics and Public Policy* **1** 6–18.

GELMAN, A. and RUBIN, D. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–511.

ICF INTERNATIONAL INC. (2014). Guidelines for the MEASURE DHS Phase III Main Survey Report. Technical report. Available at http://dhsprogram.com/pubs/pdf/DHSM6/Final_Report_Tab_Plan_24Oct2014_DHSM6.pdf.

OESTERGAARD, M. Z., ALKEMA, L. and LAWN, J. E. (2013). Millennium Development Goals national targets are moving targets and the results will not be known until well after the deadline of 2015. *Int. J. Epidemiol.* **42** 645–647.

PLUMMER, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the* 3*rd International Workshop on Distributed Statistical Computing* (*DSC* 2003), *March* 20–22, *Vienna*, *Austria*. Available at http://mcmc-jags.sourceforge.net/. ISSN 1609-395X.

UNAIDS (2013). *Global Report*: *UNAIDS Report on the Global AIDS Epidemic* 2013. Geneva: Joint United Nations Programme on HIV/AIDS.

UNITED NATIONS POPULATION DIVISION (2013). *World Population Prospects*. *The* 2012 *Revision*. United Nations publication.

WHO, UNICEF, UNFPA, THE WORLD BANK and THE UNITED NATIONS POPULATION DIVISION (2014). Trends in maternal mortality 1990–2013: Estimates developed by WHO, UNICEF, UNFPA, The World Bank and the United Nations Population Division. ISBN 978 92 4 150722 6.

WHO, UNICEF, UNFPA, THE WORLD BANK and THE UNITED NATIONS POPULATION DIVISION (2015). Trends in maternal mortality 1990–2015: Estimates developed by WHO, UNICEF, UNFPA. The World Bank and the United Nations Population Division.

WILMOTH, J. R., MIZOGUCHI, N., OESTERGAARD, M. Z., SAY, L., MATHERS, C. D., ZUREICK-BROWN, S., INOUE, M. and CHOU, D. (2012). A new method for deriving global estimates of maternal mortality. *Statistics*, *Politics*, *and Policy* **3**.

WORLD HEALTH ORGANIZATION (2010). *International Statistical Classification of Diseases and Related Health Problems*, *Tenth Revision*: *Instruction Manual*. World Health Organization, Geneva.

WORLD HEALTH ORGANIZATION (2014). *Life Tables for WHO Member States* 1990–2012. World Health Organization, Geneva.

WORLD HEALTH ORGANIZATION (2015). Strategies toward ending preventable maternal mortality (EPMM). Technical report. Available at http://apps.who.int/iris/bitstream/10665/153544/1/9789241508483_eng.pdf?ua=1.

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1275–1308
DOI: 10.1214/17-AOAS1025

MAXIMUM LIKELIHOOD FEATURES FOR GENERATIVE IMAGE MODELS[1]

BY LO-BIN CHANG[*], ERAN BORENSTEIN[†], WEI ZHANG[‡] AND STUART GEMAN[§]

Ohio State University[*], Amazon[†], Smartleaf[‡] and Brown University[§]

Most approaches to computer vision can be thought of as lying somewhere on a continuum between generative and discriminative. Although each approach has had its successes, recent advances have favored discriminative methods, most notably the convolutional neural network. Still, there is some doubt about whether this approach will scale to a human-level performance given the numbers of samples that are needed to train state-of-the-art systems. Here, we focus on the generative or Bayesian approach, which is more model based and, in theory, more efficient. Challenges include latent-variable modeling, computationally efficient inference, and data modeling. We restrict ourselves to the problem of data modeling, which is possibly the most daunting, and specifically to the generative modeling of image patches. We formulate a new approach, which can be broadly characterized as an application of "conditional modeling," designed to sidestep the high-dimensionality and complexity of image data. A series of experiments, learning appearance models for faces and parts of faces, illustrates the flexibility and effectiveness of the approach.

REFERENCES

AGARWAL, S., AWAN, A. and ROTH, D. (2004). Learning to detect objects in images via a sparse part-based representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 1475–1490.

AHARON, M., ELAD, M. and BRUCKSTEIN, A. M. (2006). The KSVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Trans. Signal Process.* **54** 4311–4322.

ALLASSONNIÈRE, S., AMIT, Y. and TROUVÉ, A. (2007). Towards a coherent statistical framework for dense deformable template estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 3–29. MR2301497

AMIT, Y., GEMAN, D. and FAN, X. (2004). A coarse-to-fine strategy for multiclass shape detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 1606–1621.

AMIT, Y. and TROUVÉ, A. (2006). *Generative Models for Labeling Multi-Object Configurations in Images* 362–381. Springer Berlin, Heidelberg.

AMIT, Y. and TROUVÉ, A. (2007). POP: Patchwork of parts models for object recognition. *Int. J. Comput. Vis.* **75** 267–282.

BLANCHARD, G. and GEMAN, D. (2005). Hierarchical testing designs for pattern recognition. *Ann. Statist.* **33** 1155–1202. MR2195632

BORENSTEIN, E. and ULLMAN, S. (2002). Class-specific, top-down segmentation. In *ECCV. LNCS* **2353** 109–122.

Key words and phrases. Computer vision, image models, appearance models, generative models, conditional modeling, sufficiency, features.

CHUNG, K. L. (2001). *A Course in Probability Theory*, 3rd ed. Academic Press, Inc., San Diego, CA. MR1796326

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

DIACONIS, P. and FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12** 793–815. MR0751274

DÜMBGEN, L. and DEL CONTE-ZERIAL, P. (2013). On low-dimensional projections of high-dimensional distributions. In *From Probability to Statistics and Back: High-Dimensional Models and Processes. Inst. Math. Stat. (IMS) Collect.* **9** 91–104. IMS, Beachwood, OH. MR3186751

FELDMAN, T. and YOUNES, L. (2006). Homeostatic image perception: An artificial system. *Comput. Vis. Image Underst.* **102** 70–80.

FELZENSZWALB, P. (2013). A stochastic grammar for natural shapes. In *Shape Perception in Human and Computer Vision* (S. J. Dickinson and Z. Pizlo, eds.) 299–310. Springer, London.

FELZENSZWALB, P. F., GIRSHICK, R. B., MCALLESTER, D. and RAMANAN, D. (2010). Object detection with discriminatively trained part based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **32** 1627–1645.

FERGUS, R., PERONA, P. and ZISSERMAN, A. (2003). Object class recognition by unsupervised scale-invariant learning. *CVPR* **2** 264–271.

FREY, B. J. (2003). Transformation-invariant clustering using the EM algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* **25** 1–17.

FREY, B. J. and JOJIC, N. (1999). Transformed component analysis: Joint estimation of spatial transformations and image components. In *International Conference on Computer Vision* **2** 1190.

HEISELE, B., SERRE, T. and POGGIO, T. (2007). A component-based framework for face detection and identification. *Int. J. Comput. Vis.* **74** 167–181.

HEISELE, B., SERRE, T., PONTIL, M., VETTER, T. and POGGIO, T. (2001). Categorization by learning and combining object parts. In *NIPS*.

HINTON, G. E. (1999). Products of experts. In *Int. Conf. on Art. Neur. Netw. (ICANN)* **1** 1–6.

JIN, Y. and GEMAN, S. (2006). Context and hierarchy in a probabilistic image model. In *CVPR* 2145–2152.

KANNAN, A., JOJIC, N. and FREY, B. (2002). Fast transformation invariant factor analysis. In *Advances in Neural Information Processing Systems* **15**.

LEE, H., BATTLE, A., RAINA, R. and NG, A. Y. (2007). Efficient sparse coding algorithms. *Adv. Neural Inf. Process. Syst.* **19** 801–808.

LEIBE, B. and SCHIELE, B. (2003). Interleaved object categorization and segmentation. In *Proceedings of British Machine Vision Conference (BMVC)*.

MAIRAL, J., BACH, F., PONCE, J. and SAPIRO, G. (2009). Online dictionary learning for sparse coding. In *Proceedings of the 26th International Conference on Machine Learning*.

OLSHAUSEN, B. A. and FIELD, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vis. Res.* **37** 3311–3325.

OMMER, B. and BUHMANN, J. M. (2006). Learning compositional categorization models. In *ECCV*.

PAPANDREOU, G., CHEN, L.-C. and YUILLE, A. (2014). Modeling image patches with a generic dictionary of mini-epitomes. In *Proc. IEEE Int. Conf. on Comp. Vision and Pat. Rec. (CVPR)*.

RAJAGOPALAN, A. N., CHELLAPPA, R. and KOTERBA, N. T. (2005). Background learning for robust face recognition with PCA in the presence of clutter. *IEEE Trans. Image Process.* **14** 832–843.

REID, N. (1995). The roles of conditioning in inference. *Statist. Sci.* **10** 138–157, 173–189, 193–196. With comments by V. P. Godambe, Bruce G. Lindsay and Bing Li, Peter McCullagh, George Casella, Thomas J. DiCiccio and Martin T. Wells, A. P. Dawid and C. Goutis and Thomas Severini, with a rejoinder by the author. MR1368097

ROTH, S. and BLACK, M. J. (2009). Fields of experts. *Int. J. Comput. Vis.* **82** 205–229.

SABUNCU, M. R., BALCI, S. K. and GOLLAND, P. (2008). Discovering modes of an image population through mixture modeling. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention* (*MICCAI*). *LNCS* **5242** 381–389.

SALI, E. and ULLMAN, S. (1999). Combining class-specific fragments for object classification. In *Proc*. 10*th British Machine Vision Conference* **1** 203–213.

SI, Z. and ZHU, S.-C. (2012). Learning hybrid image templates (HIT) by information projection. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **34** 1354–1367.

ULLMAN, S., SALI, E. and VIDAL-NIQUET, M. (2001). A fragment-based approach to object representation and classification. In *International Workshop on Visual Form* 85–100.

ULLMAN, S., VIDAL-NAQUET, M. and SALI, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat*. *Neurosci*. **5** 682–687.

WEBER, M., WELLING, M. and PERONA, P. (2000). Unsupervised learning of models for recognition. In *Proc*. *Sixth European Conf*. *Computer Vision* 18–22.

WELLING, M., HINTON, G. E. and OSINDERO, S. (2003). Learning sparse topographic representations with products of student-t distributions. In *Adv. in Neur. Inf. Proc. Sys*. (*NIPS*) **15** 1359–1366.

YUILLE, A. (2011). Towards a theory of compositional learning and encoding of objects. In *Computational Methods for the Innovative Design of Electrical Devices'*11 1448–1455.

ZHU, L., CHEN, Y. and YUILLE, A. (2009). Unsupervised learning of probabilistic grammar-Markov models for object categories. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **31** 114–128.

ZHU, S.-C. and MUMFORD, D. (1997). Prior learning and Gibbs reaction-diffusion. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **19** 1236–1250.

ZHU, S.-C. and MUMFORD, D. (2006). A stochastic grammar of images. In *Foundations and Trends in Computer Graphics and Vision* 259–362.

ZHU, S.-C., WU, Y. and MUMFORD, D. (1998). Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *Int*. *J*. *Comput*. *Vis*. **27**.

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1309–1348
DOI: 10.1214/17-AOAS1027

BAYESIAN ESTIMATES OF ASTRONOMICAL TIME DELAYS BETWEEN GRAVITATIONALLY LENSED STOCHASTIC LIGHT CURVES[1]

BY HYUNGSUK TAK[*,2,3], KAISEY MANDEL[†,4], DAVID A. VAN DYK[‡,6], VINAY L. KASHYAP[‡,5], XIAO-LI MENG[†,2] AND ANETA SIEMIGINOWSKA[†,5]

Statistical and Applied Mathematical Sciences Institute[*], Harvard University[†] and Imperial College London[‡]

The gravitational field of a galaxy can act as a lens and deflect the light emitted by a more distant object such as a quasar. Strong gravitational lensing causes multiple images of the same quasar to appear in the sky. Since the light in each gravitationally lensed image traverses a different path length from the quasar to the Earth, fluctuations in the source brightness are observed in the several images at different times. The time delay between these fluctuations can be used to constrain cosmological parameters and can be inferred from the time series of brightness data or light curves of each image. To estimate the time delay, we construct a model based on a state-space representation for irregularly observed time series generated by a latent continuous-time Ornstein–Uhlenbeck process. We account for microlensing, an additional source of independent long-term extrinsic variability, via a polynomial regression. Our Bayesian strategy adopts a Metropolis–Hastings within Gibbs sampler. We improve the sampler by using an ancillarity-sufficiency interweaving strategy and adaptive Markov chain Monte Carlo. We introduce a profile likelihood of the time delay as an approximation of its marginal posterior distribution. The Bayesian and profile likelihood approaches complement each other, producing almost identical results; the Bayesian method is more principled but the profile likelihood is simpler to implement. We demonstrate our estimation strategy using simulated data of doubly- and quadruply-lensed quasars, and observed data from quasars *Q0957+561* and *J1029+2623*.

REFERENCES

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2015). Overall objective priors. *Bayesian Anal.* **10** 189–221. MR3420902

BERGER, J. O., LISEO, B. and WOLPERT, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.* **14** 1–28. MR1702200

BERK, D. E. V., WILHITE, B. C., KRON, R. G., ANDERSON, S. F., BRUNNER, R. J., HALL, P. B., IVEZIĆ, Ž., RICHARDS, G. T., SCHNEIDER, D. P., YORK, D. G., BRINKMANN, J. V., LAMB, D. Q., NICHOL, R. C. and SCHLEGEL, D. J. (2004). The ensemble photometric variability of ∼25,000 quasars in the sloan digital sky survey. *Astrophys. J.* **601** 692.

BLANDFORD, R. and NARAYAN, R. (1992). Cosmological applications of gravitational lensing. *Annu. Rev. Astron. Astrophys.* **30** 311–358.

Key words and phrases. Gravitational lensing, microlensing, Ornstein–Uhlenbeck process, Gibbs sampler, profile likelihood, ancillarity-sufficiency interweaving strategy, adaptive MCMC, Q0957+561, J1029+2623, LSST, quasar.

BROOKS, S. P., MORGAN, B. J., RIDOUT, M. S. and PACK, S. (1997). Finite mixture models for proportions. *Biometrics* **53** 1097–1115.

BROOKS, S., GELMAN, A., JONES, G. L. and MENG, X.-L., eds. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press, Boca Raton, FL. MR2742422

CHANG, K. and REFSDAL, S. (1979). Flux variations of QSO 0957+561 A, B and image splitting by stars near the light path. *Nature* **282** 561–564.

COURBIN, F., CHANTRY, V., REVAZ, Y., SLUSE, D., FAURE, C., TEWES, M., EULAERS, E., KOLEVA, M., ASFANDIYAROV, I., DYE, S., MAGAIN, P., VAN WINCKEL, H., COLES, J., SAHA, P., IBRAHIMOV, M. and MEYLAN, G. (2013). COSMOGRAIL: The COSmological MOnitoring of GRAvItational lenses IX. time delays, lens dynamics and baryonic fraction in He 0435-1223. *Astron. Astrophys.* **536** A53.

DAVISON, A. C. (2003). *Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **11**. Cambridge Univ. Press, Cambridge. MR1998913

DOBLER, G., FASSNACHT, C., TREU, T., MARSHALL, P. J., LIAO, K., HOJJATI, A., LINDER, E. and RUMBAUGH, N. (2015). Strong lens time delay challenge. I. Experimental design. *Astrophys. J.* **799** 168.

FASSNACHT, C., PEARSON, T., READHEAD, A., BROWNE, I., KOOPMANS, L., MYERS, S. and WILKINSON, P. (1999). A determination of h_o with the CLASS gravitational lens B1608+656. I. time delay measurements with the VLA. *Astrophys. J.* **527** 498.

FISCHER, P., BERNSTEIN, G., RHEE, G. and TYSON, J. A. (1997). The mass distribution of the cluster Q0957+561 from gravitational lensing. *Astron. J.* **113** 521.

FOHLMEISTER, J., KOCHANEK, C. S., FALCO, E. E., WAMBSGANSS, J., OGURI, M. and DAI, X. (2013). A two-year time delay for the lensed quasar SDSS J1029+ 2623. *Astrophys. J.* **764** 186.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. CRC Press, Boca Raton, FL. MR3235677

HAINLINE, L. J., MORGAN, C. W., BEACH, J. N., KOCHANEK, C., HARRIS, H. C., TILLEMAN, T., FADELY, R., FALCO, E. E. and LE, T. X. (2012). A new microlensing event in the doubly imaged quasar Q 0957+561. *Astrophys. J.* **744** 104.

HARVA, M. and RAYCHAUDHURY, S. (2006). *Bayesian Estimation of Time Delays Between Unevenly Sampled Signals*. IEEE, New York.

HOJJATI, A., KIM, A. G. and LINDER, E. V. (2013). Robust strong lensing time delay estimation. *Phys. Rev. D* **87** 123512.

INADA, N., OGURI, M., MOROKUMA, T., DOI, M., YASUDA, N., BECKER, R. H., RICHARDS, G. T., KOCHANEK, C. S., KAYO, I., KONISHI, K. et al. (2006). SDSS J1029+2623: A gravitationally lensed quasar with an image separation of 225. *Astrophys. J. Lett.* **653** L97.

KELLY, B. C., BECHTOLD, J. and SIEMIGINOWSKA, A. (2009). Are the variations in quasar optical flux driven by thermal fluctuations? *Astrophys. J.* **698** 895.

KOCHANEK, C., MORGAN, N., FALCO, E., MCLEOD, B., WINN, J., DEMBICKY, J. and KETZEBACK, B. (2006). The time delays of gravitational lens He 0435–1223: An early-type galaxy with a rising rotation curve. *Astrophys. J.* **640** 47.

KOZŁOWSKI, S. and KOCHANEK, C. S. (2009). Discovery of 5000 active galactic nuclei behind the magellanic clouds. *Astrophys. J.* **701** 508.

KOZŁOWSKI, S., KOCHANEK, C. S., UDALSKI, A., SOSZYŃSKI, I., SZYMAŃSKI, M., KUBIAK, M., PIETRZYŃSKI, G., SZEWCZYK, O., ULACZYK, K. and POLESKI, R. (2010). Quantifying quasar variability as part of a general approach to classifying continuously varying sources. *Astrophys. J.* **708** 927.

KUMAR, S. R., STALIN, C. and PRABHU, T. (2014). H_0 from 11 well measured time-delay lenses. *Astron. Astrophys.* **580** A38.

LIAO, K., TREU, T., MARSHALL, P., FASSNACHT, C. D., RUMBAUGH, N., DOBLER, G., AGHAMOUSA, A., BONVIN, V., COURBIN, F., HOJJATI, A., JACKSON, N., KASHYAP, V., RATHNA KUMAR, S., LINDER, E., MANDEL, K., MENG, X. L., MEYLAN, G., MOUSTAKAS, L. A., PRABHU, T. P., ROMERO-WOLF, A., SHAFIELOO, A., SIEMIGINOWSKA, A., STALIN, C. S., TAK, H., TEWES, M. and VAN DYK, D. (2015). Strong lens time delay challenge: II. Results of TDC1. *Astrophys. J.* **800** 11.

LINDER, E. V. (2011). Lensing time delays and cosmological complementarity. *Phys. Rev. D* **84** 123529.

LIU, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR2401592

LSST SCIENCE COLLABORATION (2009). LSST Science Book, Version 2.0. Available at arXiv:0912.0201.

MACLEOD, C., IVEZIĆ, Ž., KOCHANEK, C., KOZŁOWSKI, S., KELLY, B., BULLOCK, E., KIMBALL, A., SESAR, B., WESTMAN, D., BROOKS, K., GIBSON, R., BECKER, A. C. and DE VRIES, W. H. (2010). Modeling the time variability of SDSS stripe 82 quasars as a damped random walk. *Astrophys. J.* **721** 1014.

MORGAN, C. W., HAINLINE, L. J., CHEN, B., TEWES, M., KOCHANEK, C. S., DAI, X., KOZLOWSKI, S., BLACKBURNE, J. A., MOSQUERA, A. M., CHARTAS, G., COURBIN, F. and MEYLAN, G. (2012). Further evidence that quasar X-ray emitting regions are compact: X-ray and optical microlensing in the lensed quasar q J0158-4325. *Astrophys. J.* **756** 52.

MOSQUERA, A. M. and KOCHANEK, C. S. (2011). The microlensing properties of a sample of 87 lensed quasars. *Astrophys. J.* **738** 96.

MUNOZ, J., FALCO, E., KOCHANEK, C., LEHÁR, J., MCLEOD, B., IMPEY, C., RIX, H.-W. and PENG, C. (1998). The CASTLES project. *Astrophys. Space Sci.* **263** 51–54.

OGURI, M. and MARSHALL, P. J. (2010). Gravitationally lensed quasars and supernovae in future wide-field optical imaging surveys. *Mon. Not. R. Astron. Soc.* **405** 2579–2593.

OSCOZ, A., MEDIAVILLA, E., GOICOECHEA, L. J., SERRA-RICART, M. and BUITRAGO, J. (1997). Time delay of QSO 0957+561 and cosmological implications. *Astrophys. J. Letters* **479** L89.

OSCOZ, A., ALCALDE, D., SERRA-RICART, M., MEDIAVILLA, E., ABAJAS, C., BARRENA, R., LICANDRO, J., MOTTA, V. and MUNOZ, J. (2001). Time delay in QSO 0957+561 from 1984-1999 optical data. *Astrophys. J.* **552** 81.

PELT, J., HOFF, W., KAYSER, R., REFSDAL, S. and SCHRAMM, T. (1994). Time delay controversy on QSO 0957+ 561 not yet decided. *Astron. Astrophys.* **286** 775–785.

PELT, J., KAYSER, R., REFSDAL, S. and SCHRAMM, T. (1996). The light curve and the time delay of QSO 0957+561. *Astron. Astrophys.* **305** 97–106.

R CORE TEAM (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

REFSDAL, S. (1964). The gravitational lens effect. *Mon. Not. R. Astron. Soc.* **128** 295–306. MR0175607

ROBERTS, G. O. and ROSENTHAL, J. S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.* **44** 458–475. MR2340211

SCHNEIDER, P., EHLERS, J. and FALCO, E. (1992). *Gravitational Lenses*. Springer, Berlin.

SCHNEIDER, P., WAMBSGANSS, J. and KOCHANEK, C. S. (2006). *Gravitational Lensing*: *Strong, Weak and Micro*. Springer, Berlin.

SERRA-RICART, M., OSCOZ, A., SANCHÍS, T., MEDIAVILLA, E., GOICOECHEA, L. J., LICANDRO, J., ALCALDE, D. and GIL-MERINO, R. (1999). BVRI photometry of QSO 0957+561A, B: Observations, new reduction method, and time delay. *Astrophys. J.* **526** 40.

SHALYAPIN, V., GOICOECHEA, L. and GIL-MERINO, R. (2014). A 5.5-year robotic optical monitoring of Q0957+561: Substructure in a non-local cD galaxy. *Astron. Astrophys.* **540** A132.

SUYU, S., AUGER, M., HILBERT, S., MARSHALL, P., TEWES, M., TREU, T., FASSNACHT, C., KOOPMANS, L., SLUSE, D., BLANDFORD, R., COURBIN, F. and MEYLAN, G. (2013). Two accurate time-delay distances from strong lensing: Implications for cosmology. *Astrophys. J.* **766** 70.

TAK, H., KELLY, J. and MORRIS, C. N. (2017). Rgbp: An R package for Gaussian, Poisson, and Binomial Random Effects Models with Frequency Coverage Evaluations. *J. Stat. Softw.* **78** 1–33.

TAK, H., MANDEL, K., VAN DYK, D. A., KASHYAP, V. L., MENG, X.-L and SIEMIGINOWSKA, A. (2017). Supplement to "Bayesian estimates of astronomical time delays between gravitationally lensed stochastic light curves." DOI:10.1214/17-AOAS1027SUPP.

TEWES, M., COURBIN, F. and MEYLAN, G. (2013). COSMOGRAIL: The COSmological MOnitoring of GRAvItational lenses XI. techniques for time delay measurement in presence of microlensing. *Astrophys. J.* **605** 58.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762. MR1329166

TREU, T. and MARSHALL, P. J. (2016). Time delay cosmography. *Astron. Astrophys. Rev.* **24** 11.

UHLENBECK, G. E. and ORNSTEIN, L. S. (1930). On the theory of the Brownian motion. *Phys. Rev.* **36** 823–841.

VAN DYK, D. A. and MENG, X.-L. (2001). The art of data augmentation. *J. Comput. Graph. Statist.* **10** 1–111. MR1936358

WALSH, D., CARSWELL, R. and WEYMANN, R. (1979). 0957+561 A, B- twin quasistellar objects or gravitational lens. *Nature* **279** 381–384.

YU, Y. and MENG, X.-L. (2011). To center or not to center: That is not the question—an ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *J. Comput. Graph. Statist.* **20** 531–570. MR2878987

ZU, Y., KOCHANEK, C., KOZŁOWSKI, S. and UDALSKI, A. (2013). Is quasar optical variability a damped random walk? *Astrophys. J.* **765** 106.

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1349–1374
DOI: 10.1214/17-AOAS1028

COMPARING HEALTHCARE UTILIZATION PATTERNS VIA GLOBAL DIFFERENCES IN THE ENDORSEMENT OF CURRENT PROCEDURAL TERMINOLOGY CODES[1]

BY XU SHI*, HRISTINA PASHOVA† AND PATRICK J. HEAGERTY*

*University of Washington** *and Axio Research*†

The linkage of electronic medical records (EMR) across clinics, hospitals, and healthcare systems is opening new opportunities to evaluate factors associated with both individual treatment benefit and potential harm. For example, the FDA Sentinel initiative seeks to create a surveillance network with over 100 million patient lives (Behrman et al. [*N. Engl. J. Med.* **364** (2011) 498–499]), while PCORnet has created multiple networks that include linked electronic medical records from geographic regions such as entire cities or states, with the ultimate goal of facilitating comparative effectiveness research (Collins et al. [*Journal of the American Medical Informatics Association* **4** (2014) 576–577]). However, one key challenge to the use of electronically assembled cohorts is the potential for variation in both the choice of specific healthcare procedures and coding practices due to differences in patient populations and/or financial incentives within care delivery networks. In order to explore variation in patient care or procedure coding, we review and develop statistical methods that can permit testing and estimation of subgroup differences in code assignments. We focus on Current Procedural Terminology (CPT) codes which are used in a standardized fashion to capture patient treatment details and to record medical histories, but the methods we develop can be used for any structured EMR data. We specifically study testing procedures that can be valid for comparing both rare and common counts as routinely encountered with medical procedure codes, and we transfer methods from studies of genetic association. Hierarchical structure in terms of both thematically grouped medical codes and provider-level clustering adds unique complexity to the analysis of EMR data. We detail penalized regression methods unifying estimation and inference to leverage the hierarchical structure and stabilize rate ratio estimates for rare procedures. We also expand inference methods to account for potential within provider correlation of patient utilization. We illustrate methods comparing the endorsement of CPT codes for subjects enrolled in a back pain cohort study where interest is in the differences across recruitment centers in the use of CPT codes (Jarvik [*BMC Musculoskelet Disord.* **13** (2012)]).

REFERENCES

BASU, S. and PAN, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology* **35** 606–619.

Key words and phrases. Electronic medical records, hierarchical structure, dynamic graphics.

BEHRMAN, R. E., BENNER, J. S., BROWN, J. S., MCCLELLAN, M., WOODCOCK, J. and PLATT, R. (2011). Developing the Sentinel System—A national resource for evidence development. *N. Engl. J. Med.* **364** 498–499.

BENTLEY, P. N., WILSON, A. G., DERWIN, M. E., SCODELLARO, R. and JACKSON, R. E. (2002). Reliability of assigning correct current procedural terminology—4 E/M codes. *Ann. Emerg. Med.* **40** 269–274.

BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549

BULL, S. B. (1998). Regression models for multiple outcomes in large epidemiologic studies. *Stat. Med.* **17** 2179–2197.

CHAPMAN, J. and WHITTAKER, J. (2008). Analysis of multiple SNPs in a candidate gene or region. *Genetic Epidemiology* **32** 560–566.

COLLINS, F. S., HUDSON, K. L., BRIGGS, J. P. and LAUER, M. S. (2014). PCORnet: Turning a dream into reality. *Journal of the American Medical Informatics Association* **4** 576–577.

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR2049007

HOERL, A. and KENNARD, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

HOLT, J., WARSY, A. and WRIGHT, P. (2010). Medical decision making: Guide to improved CPT coding. *South. Med. J.* **103** 316–322.

JARVIK, J. G. (2012). Study protocol: The back pain outcomes using longitudinal data (BOLD) registry. *BMC Musculoskelet Disord.* **13** 64.

KING, M. S., LIPSKY, M. S. and SHARP, L. (2002). Expert agreement in Current Procedural Terminology evaluation and management coding. *Arch. Intern. Med.* **162** 316–320.

KING, M. S., SHARP, L. and LIPSKY, M. S. (2001). Accuracy of CPT evaluation and management coding by family physicians. *J. Am. Board Fam. Pract.* **14** 184–192.

LEE, S., EMOND, M. J., BASHED, M. J., BARNES, K. C., RIEDER, M. J., NICKERSON, D. A., NHLBI GO EXOME SEQUENCING PROJECT—ESP LUNG PROJECT TEAM, CHRISTIANI, D. C., WURZEL, M. M. and LIN, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91** 224–237.

MADSEN, B. E. and BROWNING, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5** e1000384.

MIGLIORETTI, D. L. and HEAGERTY, P. J. (2004). Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics* **5** 381–398.

MIGLIORETTI, D. L. and HEAGERTY, P. J. (2007). Marginal modeling of nonnested multilevel data using standard software. *Am. J. Epidemiol.* **165** 453–463.

MORGENTHALER, S. and THILLY, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutat. Res.* **615** 28–56.

MORRIS, A. P. and ZEGGINI, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology* **34** 188–193.

PAN, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology* **33** 487–507.

PRZYBOROWSKI, J. and WILENSKI, H. (1940). Homogeneity of results in testing samples from Poisson series with an application to testing clover seed for dodder. *Biometrika* **31** 313–323. MR0002070

QI, Y., WEEKS, D. E., TIWARI, H. K., YI, N., ZHANG, K., GAO, G., LIN, W., LOU, X., CHEN, W. and LIU, W. (2015). Rare-variant kernel machine test for longitudinal data from population and family samples. *Hum. Hered.* **80** 126–138.

ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, G. K. (2010). EdgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 515–524.

SHI, X., PASHOVA, H. and HEAGERTY, P. J. (2017). Supplement to "Comparing healthcare utilization patterns via global differences in the endorsement of current procedural terminology codes." DOI:10.1214/17-AOAS1028SUPPA, DOI:10.1214/17-AOAS1028SUPPB, DOI:10.1214/17-AOAS1028SUPPC, DOI:10.1214/17-AOAS1028SUPPD.

WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89** 82–93.

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1375–1402
DOI: 10.1214/17-AOAS1029

GAUSSIAN PROCESS FRAMEWORK FOR TEMPORAL DEPENDENCE AND DISCREPANCY FUNCTIONS IN RICKER-TYPE POPULATION GROWTH MODELS

BY MARCELO HARTMANN[*,1,2], GEOFFREY R. HOSACK[†], RICHARD M. HILLARY[†] AND JARNO VANHATALO[*,1,2]

University of Helsinki[*] and CSIRO Marine Laboratories[†]

Density dependent population growth functions are of central importance to population dynamics modelling because they describe the theoretical rate of recruitment of new individuals to a natural population. Traditionally, these functions are described with a fixed functional form with temporally constant parameters and without species interactions. The Ricker stock-recruitment model is one such function that is commonly used in fisheries stock assessment. In recent years, there has been increasing interest in semiparametric and temporally varying population growth models. The former are related to the general statistical approach of using semiparametric discrepancy functions, such as Gaussian processes (GP), to model deviations around the expected parametric function. In the latter, the reproductive rate, which is a key parameter describing the population growth rate, is assumed to vary in time. In this work, we introduce how these existing Ricker population growth models can be formulated under the same statistical approach of hierarchical GP models. We also show how the time invariant semiparametric approach can be extended and combined with the time varying reproductive rate using a GP model. Then we extend these models to the multispecies setting by incorporating cross-covariances among species with a continuous time covariance structure using the linear model of coregionalization. As a case study, we examine the productivity of three Pacific salmon populations. We compare the alternative Ricker population growth functions using model posterior probabilities and leave-one-out cross validation predictive densities. Our results show substantial temporal variation in maximum reproductive rates and reveal temporal dependence among the species, which have direct management implications. However, our results do not support inclusion of semiparametric discrepancy function and they suggest that the semiparametric discrepancy functions may lead to challenges in parameter identifiability more generally.

REFERENCES

BARNARD, J., MCCULLOCH, R. and MENG, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist. Sinica* **10** 1281–1311. MR1804544

BRADFORD, M. J. (1995). Comparative review of Pacific salmon survival rates. *Can. J. Fish. Aquat. Sci.* **52** 1327–1338.

Key words and phrases. Model evidence, marginal likelihood, population growth, fisheries, density dependence, temporal dependence, interspecific dependence.

BRADFORD, M. J., LOVY, J. and PATTERSON, D. A. (2010). Infection of gill and kidney of Fraser River sockeye salmon, Oncorhynchus nerka (Walbaum), by Parvicapsula minibicornis and its effect on host physiology. *J. Fish Dis.* **33** 769–779.

BRÄNNSTRON, Å. and SUMPTER, D. J. T. (2005). The role of competition and clustering in population dynamics. *Proc. R. Soc. Lond., B Biol. Sci.* **272** 2065–2072.

BRYNJARSDÓTTIR, J. and O'HAGAN, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Probl.* **30** 114007, 24. MR3274591

BUCKLAND, S. T., NEWMAN, K. B., FERNÁNDEZ, C., THOMAS, L. and HARWOOD, J. (2007). Embedding population dynamics models in inference. *Statist. Sci.* **22** 44–58. MR2408660

BURGNER, R. L. (1991). Life history of sockeye salmon (Oncorhynchus nerka). In *Pacific Salmon Life Histories* 3–117. UBC Press, Vancouver, BC.

CADIGAN, N. G. (2013). Fitting a non-parametric stock-recruitment model in R that is useful for deriving MSY reference points and accounting for model uncertainty. *ICES J. Mar. Sci.* **70** 56–67.

CHEN, D. G. and IRVINE, J. R. (2001). A semiparametric model to examine stock-recruitment relationships incorporating environmental data. *Can. J. Fish. Aquat. Sci.* **58** 1178.

CHEN, D. G. and WARE, D. M. (1999). A neural network model for forecasting fish stock recruitment. *Can. J. Fish. Aquat. Sci.* **56** 2385–2396.

COOKE, S. J., HINCH, S. G., ANTHONY, P. F., LAPOINTE, M. F., JONES, S. R. M., MACDONALD, J. S., PATTERSON, D. A., HEALEY, M. C. and KRAAK, G. V. D. (2004). Abnormal migration timing and high en route mortality of sockeye salmon in the Fraser River, British Columbia. *Fisheries* **29** 22–33.

DFO (2016). Department of Fisheries and Oceans. Salmonid Enhancement Program, Weaver Creek Spawning Channel. Unpublished data.

DORNER, B., PETERMAN, R. M. and HAESEKER, S. L. (2008). Historical trends in productivity of 120 Pacific pink, chum, and sockeye salmon stocks reconstructed by using a Kalman filter. *Can. J. Fish. Aquat. Sci.* **65** 1842–1866.

ESSINGTON, T. E., QUINN, T. P. and EWERT, V. E. (2000). Intra- and inter-specific competition and the reproductive success of sympatric Pacific salmon. *Can. J. Fish. Aquat. Sci.* **57** 205–213.

GELFAND, A. E., KIM, H.-J., SIRMANS, C. F. and BANERJEE, S. (2003). Spatial modeling with spatially varying coefficient processes. *J. Amer. Statist. Assoc.* **98** 387–396. MR1995715

GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, 4 (*PeñíScola*, 1991) 169–193. Oxford Univ. Press, New York. MR1380276

HEALEY, M. C. (1991). Life history of chinook salmon (*Oncorhynchus tshawytscha*). In *Pacific Salmon Life Histories* 311–394. UBC Press, Vancouver, BC.

HEARD, W. R. (1991). Life history of pink salmon (*Oncorhynchus gorbuscha*). In *Pacific Salmon Life Histories* 119–230. UBC Press, Vancouver, BC.

HILBORN, R. and WALTERS, C. J. (1992). *Quantitative Fisheries Stock Assessment*: *Choice*, *Dynamics and Uncertainty*. Chapman & Hall, New York.

HILLARY, R. M. (2012). Practical uses of non-parametric methods in fisheries assessment modelling. *Mar. Freshw. Res.* **63** 606.

HINCH, S. G., COOKE, S. J., FARRELL, A. P., MILLER, K. M., LAPOINTE, M. and PATTERSON, D. A. (2012). Dead fish swimming: A review of research on the early migration and high premature mortality in adult Fraser River sockeye salmon Oncorhynchus nerka. *J. Fish Biol.* **81** 576–599.

JOST, C. and ELLNER, S. P. (2000). Testing for predator dependence in predator–prey dynamics: A non-parametric approach. *Proc. R. Soc. Lond., B Biol. Sci.* **267** 1611–1620.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402

KENNEDY, M. C. and O'HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. MR1858398

KUIKKA, S., VANHATALO, J., PULKKINEN, H., MÄNTYNIEMI, S. and CORANDER, J. (2014). Experiences in Bayesian inference in Baltic salmon management. *Statist. Sci.* **29** 42–49. MR3201845

LEWIS, S. M. and RAFTERY, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *J. Amer. Statist. Assoc.* **92** 648–655. MR1467855

MÄNTYNIEMI, S., UUSITALO, L., PELTONEN, H., HAAPASAARI, P. and KUIKKA, S. (2013). Integrated, age-structured, length-based stock assessment model with uncertain process variances, structural uncertainty, and environmental covariates: Case of central Baltic herring. *Can. J. Fish. Aquat. Sci.* **1326** 1317–1326.

MÄNTYNIEMI, S., WHITLOCK, R. E., PERÄLÄ, T. A., BLOMSTEDT, P. A., VANHATALO, J. P., RICON, M. M., KUPARINEN, A. K., PULKKINEN, H. P. and KUIKKA, S. O. (2015). General state-space population dynamics model for Bayesian stock assessment. *ICES J. Mar. Sci.* **72** 2209–2222.

MARDIA, K. V. and GOODALL, C. R. (1993). Spatial–temporal analysis of multivariate environmental monitoring data. In *Multivariate Environmental Statistics*. *North-Holland Ser. Statist. Probab.* **6** 347–386. North-Holland, Amsterdam. MR1268443

MAUNDER, M. N. and DERISO, R. B. (2011). A state–space multistage life cycle model to evaluate population impacts in the presence of density dependence: Illustrated with application to delta smelt (*Hyposmesus transpacificus*). *Can. J. Fish. Aquat. Sci.* **68** 1285–1306.

MINTO, C., FLEMMING, J. M., BRITTEN, G. L. and WORM, B. (2014). Productivity dynamics of Atlantic cod. *Can. J. Fish. Aquat. Sci.* **71** 203–216.

MORITA, K., MORITA, S. H. and FUKUWAKA, M.-A. (2006). Population dynamics of Japanese pink salmon (*Oncorhynchus gorbuscha*): Are recent increases explained by hatchery programs or climatic variations? *Can. J. Fish. Aquat. Sci.* **63** 55–62.

MUNCH, S. B. and KOTTAS, A. (2009). A Bayesian modeling approach for determining productivity regimes and their characteristics. *Ecol. Appl.* **19** 527–537.

MUNCH, S. B., KOTTAS, A. and MANGEL, M. (2005). Bayesian nonparametric analysis of stock–recruitment relationships. *Can. J. Fish. Aquat. Sci.* **62** 1808–1821.

MYERS, R. A., MERTZ, G. and BRIDSON, J. (1997). Spatial scales of interannual recruitment variations of marine, anadromous, and freshwater fish. *Can. J. Fish. Aquat. Sci.* **54** 1400–1407.

NEAL, R. M. (2003). Slice sampling. *Ann. Statist.* **31** 705–767. With discussions and a rejoinder by the author. MR1994729

NEWMAN, K. B., FERNÁNDEZ, C., THOMAS, L. and BUCKLAND, S. T. (2009). Monte Carlo inference for state-space models of wild animal populations. *Biometrics* **65** 572–583. MR2751482

O'HAGAN, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliab. Eng. Syst. Saf.* **91** 1290–1300.

PERÄLÄ, T. and KUPARINEN, A. (2015). Detecting regime shifts in fish stock dynamics. *Can. J. Fish. Aquat. Sci.* **72** 1619–1628.

PETERMAN, R. M., PYPER, B. J. and GROUT, J. A. (2000). Comparison of parameter estimation methods for detecting climate-induced changes in productivity of Pacific salmon (*Oncorhynchus spp*). *Can. J. Fish. Aquat. Sci.* **57** 181–191.

PETERMAN, R. M., PYPER, B. J. and MACGREGOR, B. W. (2003). Use of the Kalman filter to reconstruct historical trends in productivity of Bristol Bay sockeye salmon (*Oncorhynchus nerka*). *Can. J. Fish. Aquat. Sci.* **60** 809–824.

PULKKINEN, H. and MÄNTYNIEMI, S. (2013). Maximum survival of eggs as the key parameter of stock-recruit meta-analysis: Accounting for parameter and structural uncertainty. *Can. J. Fish. Aquat. Sci.* **70** 527–533.

QUINN, T. J. and DERISO, R. B. (1999). *Quantitative Fish Dynamics*. Oxford Univ. Press.

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA. MR2514435

RICKER, W. E. (1954). Stock and recruitment. *J. Fish. Res. Board Can.* **5** 559–623.

RIIHIMÄKI, J. and VEHTARI, A. (2010). Gaussian processes with monotonicity information. *J. Mach. Learn. Res. Workshop Conf. Proc.* **9** 645–652. (AISTATS 2010 Proceedings).

ROSBERG, G. E., SCOTT, K. J. and RITHALER, R. (1986). Review of the International Pacific Salmon Fisheries Commission's Sockeye and Pink salmon enhancement facilities on the Fraser River. Technical Report, Department of Fisheries and Oceans, Vancouver, Canada.

ROSE, K. A., JUNIOR, J. A. C., WINEMILLER, K. O., MYERS, R. A. and HILBORN, R. (2001). Compensatory density dependence in fish populations: Importance, controversy, understanding and prognosis. *Fish Fish.* **2** 293–327.

SIMPSON, D., RUE, H., RIEBLER, A., MARTINS, T. G. and SØRBYE, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.* **32** 1–28. MR3634300

SUGENO, M. and MUNCH, S. B. (2013). A semiparametric Bayesian method for detecting Allee effects. *Ecology* **94** 1196–1204.

SUNDARARAJAN, S. and KEERTHI, S. S. (2001). Predictive approaches for choosing hyperparameters in Gaussian processes. *Neural Comput.* **13** 1103–1118.

THORSON, J. T., JENSEN, O. P. and ZIPKIN, E. F. (2014). How variable is recruitment for exploited marine fishes? A hierarchical model for testing life history theory. *Can. J. Fish. Aquat. Sci.* **983** 973–983.

THORSON, J. T., ONO, K. and MUNCH, S. B. (2014). A Bayesian approach to identifying and compensating for model misspecification in population models. *Ecology* **95** 329–341.

TOKUDA, T., GOODRICH, B., VAN MECHELEN, I., GELMAN, A. and TUERLINCKX, F. (2012). Visualizing distributions of covariance matrices. Technical report, University of Leuwen, Belgium and Columbia University, USA. Available at http://www.stat.columbia.edu/~gelman/research/unpublished/Visualization.pdf.

TRAXLER, G. and RANKIN, J. (1989). An infectious hematopoietic necrosis epizootic in sockeye salmon *Oncorhynchus nerka* in Weaver Creek spawning channel, Fraser River system, BC, Canada. *Dis. Aquat. Org.* **6** 221–226.

VANHATALO, J., RIIHIMÄKI, J., HARTIKAINEN, J., JYLÄNKI, P., TOLVANEN, V. and VEHTARI, A. (2013). GPstuff: Bayesian modeling with Gaussian processes. *J. Mach. Learn. Res.* **14** 1175–1179. MR3063621

VEHTARI, A. and OJANEN, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Stat. Surv.* **6** 142–228. MR3011074

ZENG, Z., NOWIERSKI, R. M., TAPER, M. L., DENNIS, B. and KEMP, W. P. (2010). Complex population dynamics in the real world: Modeling the influence of time-varying parameters and time lags. *Ecology* **79** 2193–2209.

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1403–1428
DOI: 10.1214/17-AOAS1031

A SEMIPARAMETRIC METHOD TO SIMULATE BIVARIATE SPACE–TIME EXTREMES[1]

BY ROMAIN CHAILAN[*,†], GWLADYS TOULEMONDE[*] AND JEAN-NOEL BACRO[*]

University of Montpellier[*] and IBM France[†]

Coastal hazards raise many concerns, as their assessment involves extremely high economic and ecological stakes. In particular, studies on rarely observed but damaging events are quite numerous. In order to anticipate upcoming events of this kind, specialists need to extrapolate the results of their studies to events that have not yet occurred. Such events might be more extreme than those already observed and could therefore severely impact the coast. It is therefore paramount to propose methodologies to simulate such extreme conditions. Parametric and nonparametric statistical methods have already been used to assess environmental extreme quantities, from univariate framework to spatial context; however, they do not generally focus on the simulation of extreme environmental scenarios. This study introduces a semiparametric approach based on the Extreme Value Theory (EVT), dedicated to the simulation of extreme space–time processes. In the proposed application context, these processes describe near-shore hydraulic conditions. They nourish coastal impact models to assess hazards along the coast. The benefit of this approach is to be able to characterise coastal hazards on an event scale, meaning we can characterise the impact both in space and through time for a given extreme event. The usefulness of this space–time extreme modelling is illustrated by a risk analysis related to the long-shore impact of extreme wave events in the Gulf of Lions.

REFERENCES

BAGNOLD, R. A. (1966). An approach to the sediment transport problem. General Physics Geological Survey, Prof. Paper.

BECHLER, A., BEL, L. and VRAC, M. (2015). Conditional simulations of the extremal t process: Application to fields of extreme precipitation. *Spat. Stat.* **12** 109–127. MR3346645

BEIRLANT, J., GOEGEBEUR, Y., TEUGELS, J. and SEGERS, J. (2004). *Statistics of Extremes. Theory and Applications*. Wiley, Chichester. MR2108013

BORTOT, P., COLES, S. and TAWN, J. (2000). The multivariate Gaussian tail model: An application to oceanographic data. *J. Roy. Statist. Soc. Ser. C* **49** 31–49. MR1817873

BOUCHETTE, F., SABATIER, F., SYLAIOS, G., MEULÉ, S., LIOU, J. L., HEURTEFEUX, H., DENAMIEL, C. and HWUNG, W. (2012). SUBDUNE tool: Quasiexplicit formulation of the water level along the shoreline. *Rev. Paralia* **12** 223–232.

BRUNEL, C., CERTAIN, R., SABATIER, F., ROBIN, N., BARUSSEAU, J. P., ALEMAN, N. and RAYNAL, O. (2014). 20th century sediment budget trends on the Western Gulf of Lions shoreface

Key words and phrases. Space-time extreme processes simulation, extreme value modelling, extreme waves, coastal hazards.

(France): An application of an integrated method for the study of sediment coastal reservoirs. *Geomorphology* **204** 625–637.

CAIRES, S., DE HAAN, L. and SMITH, R. L. (2011). On the determination of the temporal and spatial evolution of extreme events Technical Report, Deltares. Report 1202120-001-HYE-004 (for Rijkswaterstaat, Centre for Water Management).

CAMPMAS, L., BOUCHETTE, F., MEULE, S., PETITJEAN, L., SOUS, D., LIOU, J.-Y., LEROUX-MALLOUF, R., SABATIER, F. and HWUNG, H.-H. (2014). Typhoons driven morphodynamics of the Wan Tzu Liao sand barrier (Taiwan). *Coastal Eng. Proc.* **1** sediment–50.

CASTRUCCIO, S., HUSER, R. and GENTON, M. G. (2016). High-order composite likelihood inference for max-stable distributions and processes. *J. Comput. Graph. Statist.* **25** 1212–1229. MR3572037

CERC (1984). Shore Protection Manual.

CHAILAN, R. (2015). Application of Scientific Computing and Statistical Analysis to Address Coastal Hazards Ph.D. thesis University of Montpellier.

CHAILAN, R., TOULEMONDE, G., BOUCHETTE, F., LAURENT, A., SEVAULT, F. and MICHAUD, H. (2014). Spatial assessment of extreme significant waves heights in the Gulf of Lions. *Coastal Eng. Proc.* **1** management–17.

COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London. MR1932132

DAVIS, R. A., KLÜPPELBERG, C. and STEINKOHL, C. (2013a). Max-stable processes for modeling extremes observed in space and time. *J. Korean Statist. Soc.* **42** 399–414. MR3255398

DAVIS, R. A., KLÜPPELBERG, C. and STEINKOHL, C. (2013b). Statistical inference for max-stable processes in space and time. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 791–819. MR3124792

DAVISON, A. C. and HUSER, R. (2015). Statistics of extremes. *Annu. Rev. Stat. Appl.* **2** 203–235.

DAVISON, A. C., PADOAN, S. A. and RIBATET, M. (2012). Statistical modeling of spatial extremes. *Statist. Sci.* **27** 161–186. MR2963980

DE HAAN, L. (1984). A spectral representation for max-stable processes. *Ann. Probab.* **12** 1194–1204. MR0757776

DE HAAN, L. and DE RONDE, J. (1998). Sea and wind: Multivariate extremes at work. *Extremes* **1** 7–45. MR1652944

DE HAAN, L. and FERREIRA, A. (2006). *Extreme Value Theory*. Springer, New York. MR2234156

DE HAAN, L. and LIN, T. (2001). On convergence toward an extreme value distribution in $C[0, 1]$. *Ann. Probab.* **29** 467–483. MR1825160

DIEKER, A. B. and MIKOSCH, T. (2015). Exact simulation of Brown-Resnick random fields at a finite number of locations. *Extremes* **18** 301–314. MR3351818

DOMBRY, C., ENGELKE, S. and OESTING, M. (2016). Exact simulation of max-stable processes. *Biometrika* **103** 303–317. MR3509888

DOMBRY, C. and EYI-MINKO, F. (2013). Regular conditional distributions of continuous max-infinitely divisible random fields. *Electron. J. Probab.* **18** 1–21. MR3024101

DOMBRY, C., ÉYI-MINKO, F. and RIBATET, M. (2013). Conditional simulation of max-stable processes. *Biometrika* **100** 111–124. MR3034327

DOMBRY, C. and RIBATET, M. (2015). Functional regular variations, Pareto processes and peaks over threshold. *Stat. Interface* **8** 9–17. MR3320385

EASTOE, E. F. and TAWN, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 25–45. MR2662232

EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (1997). *Modelling Extremal Events. Applications of Mathematics* (*New York*) **33**. Springer, Berlin. MR1458613

EMBRECHTS, P., KOCH, E. and ROBERT, C. (2016). Space-time max-stable models with spectral separability. *Adv. in Appl. Probab.* **48** 77–97. MR3539298

ENGELKE, S., MALINOWSKI, A., KABLUCHKO, Z. and SCHLATHER, M. (2015). Estimation of Hüsler-Reiss distributions and Brown-Resnick processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 239–265. MR3299407

FERREIRA, A. and DE HAAN, L. (2014). The generalized Pareto process; with a view towards application and simulation. *Bernoulli* **20** 1717–1737. MR3263087

GOULDBY, B., MÉNDEZ, F. J., GUANCHE, Y., RUEDA, A. and MÍNGUEZ, R. (2014). A methodology for deriving extreme nearshore sea conditions for structural design and flood risk analysis. *Coastal Eng.* **88** 15–26.

GROENEWEG, J., CAIRES, S. and ROSCOE, K. (2012). Temporal and spatial evolution of extreme events. *Coastal Eng. Proc.* **1** management–9.

GUTIERREZ, B. T., PLANT, N. G., THIELER, E. R. and TURECEK, A. (2015). Using a Bayesian network to predict barrier island geomorphologic characteristics. *J. Geophys. Res., Earth Surf.* **120** 2452–2475.

HERRMANN, M. and SOMOT, S. (2008). Relevance of ERA40 dynamical downscaling for modeling deep convection in the Mediterranean Sea. *Geophys. Res. Lett.* **35**.

HERRMANN, M., SEVAULT, F., BEUVIER, J. and SOMOT, S. (2010). What induced the exceptional 2005 convection event in the northwestern Mediterranean basin? Answers from a modeling study. *J. Geophys. Res., Oceans* (1978–2012) **115**.

HUSER, R. and DAVISON, A. C. (2013). Composite likelihood estimation for the Brown-Resnick process. *Biometrika* **100** 511–518. MR3068451

HUSER, R. and DAVISON, A. C. (2014). Space-time modelling of extreme events. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 439–461. MR3164873

JONATHAN, P., EWANS, K. and RANDELL, D. (2013). Joint modelling of extreme ocean environments incorporating covariate effects. *Coastal Eng.* **79** 22–31.

LANTUÉJOUL, C. and BEL, L. (2014). Simulation conditionnelle du processus de Schlather. In 46*èmes Journées de Statistique de la SFdS*.

MICHAUD, H. (2011). Impacts des vagues sur les courants marins: Modélisation multi-échelle de la plage au plateau continental Ph.D. thesis Université Montpellier II-Sciences et Techniques du Languedoc.

MICHAUD, H., ROBIN, N., ESTOURNEL, C., MARSALEIX, P., LEREDDE, Y., CERTAIN, R. and BOUCHETTE, F. (2013). 3D hydrodynamic modeling of a microtidal barred beach (Sète, NW Mediterranean Sea) during storm conditions. In *Proc. 7th Int. Conf. on Coastal Dynamics, Arcachon France* **139** 1183–1194.

PADOAN, S. A., RIBATET, M. and SISSON, S. A. (2010). Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc.* **105** 263–277. MR2757202

RAILLARD, N., AILLIOT, P. and YAO, J. (2014). Modeling extreme values of processes observed at irregular time steps: Application to significant wave height. *Ann. Appl. Stat.* **8** 622–647. MR3192005

RASCLE, N. and ARDHUIN, F. (2013). A global wave parameter database for geophysical applications. Part 2: Model validation with improved source term parameterization. *Ocean Model.* **70** 174–188.

RIBATET, M., COOLEY, D. and DAVISON, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statist. Sinica* **22** 813–845. MR2954363

SCHLATHER, M. and TAWN, J. A. (2003). A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika* **90** 139–156. MR1966556

SHABY, B. A. (2014). The open-faced sandwich adjustment for MCMC using estimating functions. *J. Comput. Graph. Statist.* **23** 853–876. MR3224659

SHABY, B. A. and REICH, B. J. (2012). Bayesian spatial extreme value analysis to assess the changing risk of concurrent high temperatures across large portions of European cropland. *Environmetrics* **23** 638–648. MR3019056

SMITH, R. L. (1990). Max-stable processes and spatial extremes. Preprint. Univ. Surrey.
THIBAUD, E. and OPITZ, T. (2015). Efficient inference and simulation for elliptical Pareto processes. *Biometrika* **102** 855–870. MR3431558
TOLMAN, H. L. (2014). User Manual and System Documentation of WAVEWATCH III® version 4.18. Technical Report 316.
WADSWORTH, J. L. and TAWN, J. A. (2014). Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. *Biometrika* **101** 1–15. MR3180654
WANG, Y. and STOEV, S. A. (2011). Conditional sampling for spectrally discrete max-stable random fields. *Adv. in Appl. Probab.* **43** 461–483. MR2848386

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1429–1451
DOI: 10.1214/17-AOAS1032

ESTIMATING LINKS OF A NETWORK FROM TIME TO EVENT DATA

BY TSO-JUNG YEN*,[1], ZONG-RONG LEE*, YI-HAU CHEN*,[2], YU-MIN YEN† AND JING-SHIANG HWANG*

Academia Sinica and National Chengchi University†*

In this paper we develop a statistical method for identifying links of a network from time to event data. This method models the hazard function of a node conditional on event time of other nodes, parameterizing the conditional hazard function with the links of the network. It then estimates the hazard function by maximizing a pseudo partial likelihood function with parameters subject to a user-specified penalty function and additional constraints. To make such estimation robust, it adopts a pre-specified risk control on the number of false discovered links by using the Stability Selection method. Simulation study shows that under this hybrid procedure, the number of false discovered links is tightly controlled while the true links are well recovered. We apply our method to estimate a political cohesion network that drives donation behavior of 146 firms from the data collected during the 2008 Taiwanese legislative election. The results show that firms affiliated with elite organizations or firms of monopoly are more likely to diffuse donation behavior. In contrast, firms belonging to technology industry are more likely to act independently on donation.

REFERENCES

AMHED, A. and XING, E. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proc. Natl. Acad. Sci. USA* **106** 11878–11883.

ANANDKUMAR, A., TAN, V. Y. F., HUANG, F. and WILLSKY, A. S. (2012). High-dimensional structure estimation in Ising models: Local separation criterion. *Ann. Statist.* **40** 1346–1375. MR3015028

BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634

BURRIS, V. (2005). Interlocking directorates and political cohesion among corporate elites. *Am. J. Sociol.* **111** 249–283.

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067

Key words and phrases. Hazard network models, right-censored data, partial likelihood function, stability selection, political cohesion networks.

CHU, Y.-H. (1994). The realignment of business-government relations and regime transition in Taiwan. In *Business and Government in Industrialising Asia* (A. MacIntyre, ed.). Cornell Univ. Press, Ithaca, NY.

CSARDI, G. and NEPUSZ, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems* 1695.

DANAHER, P., WANG, P. and WITTEN, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 373–397. MR3164871

DANESHMAND, H., GOMEZ-RODRIGUEZ, M., SONG, L. and SCHÖLKOPF, B. (2014). Estimating diffusion networks: Recovery conditions, sample complexity & soft-thresholding algorithm. In *Proceedings of the* 31 *st International Conference on Machine Learning*. Beijing, China.

FREEMAN, L. (1977). A set of measures of centrality based on betweenness. *Sociometry* **40** 35–41.

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

GOMEZ-RODRIGUEZ, M., LESKOVEC, J. and KRAUSE, A. (2012). Inferring networks of diffusion and influence. *ACM Trans. Knowl. Discov. Data* **5**.

HOMANS, G. C. (1950). *The Human Group*. Harcourt, New York.

KHARE, K., OH, S.-Y. and RAJARATNAM, B. (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 803–825. MR3382598

LAFFERTY, J., LIU, H. and WASSERMAN, L. (2012). Sparse nonparametric graphical models. *Statist. Sci.* **27** 519–537. MR3025132

LAUMANN, E. O., MARSDEN, P. V. and PRENSKY, D. (1982). The boundary specification problem in network analysis. In *Applied Network Analysis* (R. S. Burt, M. Minor, eds.) 18–34. Sage, Beverly Hills, CA.

LEE, Z.-R. (2016). Corporate power and democracy: An analysis of business groups' campaign contributions in the 2008 legislator election. *Taiwanese Sociology* **31** 43–83.

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084

LOH, P.-L. and WAINWRIGHT, M. J. (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Ann. Statist.* **41** 3022–3049. MR3161456

MCPHERSON, M., SMITH-LOVIN, L. and COOK, J. M. (2001). Bird of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **27** 415–444.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523

MIZRUCHI, M. S. (1989). Similarity of political behavior among large American corporations. *Am. J. Sociol.* **95** 401–424.

MIZRUCHI, M. S. (1992). *The Structure of Corporate Political Action*: *Interfirm Relations and Their Consequences*. Harvard Univ. Press, Cambridge, MA.

NEWMAN, M. E. J. (2002). Assortative mixing in networks. *Phys. Rev. Lett.* **89** 208701.

NUMAZAKI, I. (1986). Network of Taiwanese big business: A preliminary analysis. *Mod. China* **12** 487–534.

PENG, J., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. MR2541591

QIU, H., HAN, F., LIU, H. and CAFFO, B. (2016). Joint estimation of multiple graphical models from high dimensional time series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 487–504. MR3454206

RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343

REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. MR3346695

ROGERS, E. M. (1995). *Diffusion of Innovations*. Free Press, New York.

SHAH, R. D. and SAMWORTH, R. J. (2013). Variable selection with error control: Another look at stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 55–80. MR3008271

STRANG, D. and SOULE, S. A. (1998). Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annu. Rev. Sociol.* **24** 265–290.

STRANG, D. and TUMA, N. B. (1993). Spatial and temporal heterogeneity in diffusion. *Am. J. Sociol.* **99** 614–639.

WATTS, D. J. and STROGATZ, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393** 440–442.

XUE, L., ZOU, H. and CAI, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *Ann. Statist.* **40** 1403–1429. MR3015030

YEN, T.-J, LEE, Z.-R, CHEN, Y.-H, YEN, Y.-M and HWANG, J.-S (2017). Supplement to "Estimating links of a network from time to event data." DOI:10.1214/17-AOAS1032SUPP.

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1452–1480
DOI: 10.1214/17-AOAS1034

A VARIATIONAL EM METHOD FOR MIXED MEMBERSHIP MODELS WITH MULTIVARIATE RANK DATA: AN ANALYSIS OF PUBLIC POLICY PREFERENCES

BY Y. SAMUEL WANG, ROSS L. MATSUEDA AND ELENA A. EROSHEVA

University of Washington

In this article, we consider modeling ranked responses from a heterogeneous population. Specifically, we analyze data from the Eurobarometer 34.1 survey regarding public policy preferences toward drugs, alcohol, and AIDS. Such policy preferences are likely to exhibit substantial differences within as well as across European nations reflecting a wide variety of cultures, political affiliations, ideological perspectives, and common practices. We use a mixed membership model to account for multiple subgroups with differing preferences and to allow each individual to possess partial membership in more than one subgroup. Previous methods for fitting mixed membership models to rank data in a univariate setting have utilized an MCMC approach and do not estimate the relative frequency of each subgroup. We propose a variational EM approach for fitting mixed membership models with multivariate rank data. Our method allows for fast approximate inference and explicitly estimates the subgroup sizes. Analyzing the Eurobarometer 34.1 data, we find interpretable subgroups which generally agree with the "left versus right" classification of political ideologies.

REFERENCES

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

AIROLDI, E. M., BLEI, D. M., EROSHEVA, E. A. and FIENBERG, S. E. (2015). Introduction to mixed membership models and methods. In *Handbook of Mixed Membership Models and Their Applications*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 3–13. CRC Press, Boca Raton, FL. MR3380022

BEAL, M. J. (2003). Variational algorithms for approximate Bayesian inference Ph.D. thesis, Univ. College, London.

BLEI, D. M. and LAFFERTY, J. D. (2005). Correlated topic models. In *Advances in Neural Information Processing Systems* **18** [*Neural Information Processing Systems*, *NIPS* 2005, *December* 5–8, 2005, *Vancouver*, *British Columbia*, *Canada*] 147–154.

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

BOTTOU, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'*2010 177–186. Physica-Verlag/Springer, Heidelberg. MR3362066

BROOKS, C. and MANZA, J. (2008). *Why Welfare States Persist*: *The Importance of Public Opinion in Democracies*. Univ. Chicago Press, Chicago, IL.

Key words and phrases. Mixed membership, rank data, variational inference, Eurobarometer, public policy.

BURSTEIN, P. (1998). Bringing the public back in: Should sociologists consider the impact of public opinion on public policy? *Social Forces* **77** 27–62.

BUSSE, L. M., ORBANZ, P. and BUHMANN, J. M. (2007). Cluster analysis of heterogeneous rank data. In *Machine Learning*, *Proceedings of the Twenty-Fourth International Conference* (*ICML* 2007), *Corvallis*, *Oregon*, *USA*, *June* 20–24, 2007 113–120.

CARON, F., TEH, Y. W. and MURPHY, T. B. (2014). Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. *Ann. Appl. Stat.* **8** 1145–1181. MR3262549

CAVADINO, M. and DIGNAN, J. (2006). Penal policy and political economy. *Criminology and Criminal Justice* **6** 435–456.

COHEN, A. and MALLOWS, C. (1983). Assessing goodness of fit of ranking models to data. *The Statistician* **32** 361–374.

EROSHEVA, E. A., FIENBERG, S. E. and JOUTARD, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *Ann. Appl. Stat.* **1** 502–537. MR2415745

EROSHEVA, E., FIENBERG, S. and LAFFERTY, J. (2004). Mixed-membership models of scientific publications. *Proc. Natl. Acad. Sci. USA* **101** 5220–5227.

GILL, J. (2008). Is partial-dimension convergence a problem for inferences from MCMC algorithms? *Polit. Anal.* **16** 153–178.

GORMLEY, I. C. and MURPHY, T. B. (2006). Analysis of Irish third-level college applications data. *J. Roy. Statist. Soc. Ser. A* **169** 361–379. MR2225548

GORMLEY, I. C. and MURPHY, T. B. (2008). A mixture of experts model for rank data with applications in election studies. *Ann. Appl. Stat.* **2** 1452–1477. MR2655667

GORMLEY, I. C. and MURPHY, T. B. (2009). A grade of membership model for rank data. *Bayesian Anal.* **4** 265–295. MR2507364

GRASMICK, H. G., DAVENPORT, E., CHAMLIN, M. B. and BURSIK, R. J. (1992). Protestant fundamentalism and the retributive doctrine of punishment. *Criminology* **30** 21–46.

GROSS, J. H. and MANRIQUE-VALLIER, D. (2015). A mixed membership approach to the assessment of political ideology from survey responses. In *Handbook of Mixed Membership Models and Their Applications*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 119–139. CRC Press, Boca Raton, FL. MR3380027

GUIVER, J. and SNELSON, E. (2009). Bayesian inference for Plackett–Luce ranking models. In *Proceedings of the* 26*th Annual International Conference on Machine Learning*, *ICML* 2009, *Montreal*, *Quebec*, *Canada*, *June* 14–18, 2009 377–384.

HUNTER, D. R. (2004). MM algorithms for generalized Bradley–Terry models. *Ann. Statist.* **32** 384–406. MR2051012

KITSCHELT, H. and REHM, P. (2014). Occupations as a site of political preference formation. *Comparative Political Studies* **47** 1670–1706.

LANGE, T., BRAUN, M. L., ROTH, V. and BUHMANN, J. M. (2002). Stability-based model selection. In *Advances in Neural Information Processing Systems* **15** [*Neural Information Processing Systems*, *NIPS* 2002, *December* 9–14, 2002, *Vancouver*, *British Columbia*, *Canada*] 617–624.

LUCE, R. D. (1977). The choice axiom after twenty years. *J. Math. Psych.* **15** 215–233. MR0462675

MARDEN, J. I. (1995). *Analyzing and Modeling Rank Data*. *Monographs on Statistics and Applied Probability* **64**. Chapman & Hall, London. MR1346107

MAYHEW, P. and VAN KESTEREN, J. (2002). Cross-national attitudes to punishment. *Changing Attitudes to Punishment* 63–92.

MEILA, M. and CHEN, H. (2010). Dirichlet process mixtures of generalized mallows models. In *UAI* 2010, *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, *Catalina Island*, *CA*, *USA*, *July* 8–11, 2010 358–367.

NOCEDAL, J. and WRIGHT, S. J. (1999). *Numerical Optimization*. *Springer Series in Operations Research*. Springer, New York. MR1713114

PLACKETT, R. L. (1975). The analysis of permutations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **24** 193–202. MR0391338

R CORE TEAM (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

REIF, K. and MELICH, A. (2001). Euro-barometer 34.1: Health Problems, Fall 1990.

ROBERTS, J. V. (2013). Public opinion and the nature of community penalties: Nternational findings. *Changing Attitudes to Punishment* 33.

SEN, A. K. (2014). *Collective Choice and Social Welfare* **11**. Elsevier, Amsterdam.

TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. MR2279480

TONRY, M. (2007). Determinants of penal policies. *Crime and Justice* **36** 1–48.

WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

WANG, C. and BLEI, D. M. (2015). A general method for robust Bayesian modeling. *ArXiv preprint. Available at arXiv*:1510.05078.

WANG, Y. S. and EROSHEVA, E. A. (2015). mixedMem: Tools for discrete multivariate mixed membership models. R package version 1.1.2. Available at https://cran.r-project.org/web/packages/mixedMem/index.html.

WANG, Y. S. and EROSHEVA, E. A. (2017). Supplement to "A variational EM method for mixed membership models with multivariate rank data: An analysis of public policy preferences." DOI:10.1214/17-AOAS1034SUPP.

ZALLER, J. (1992). *The Nature and Origins of Mass Opinion*. Cambridge Univ. Press, Cambridge, MA.

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1481–1512
DOI: 10.1214/17-AOAS1042

A NOVEL AND EFFICIENT ALGORITHM FOR DE NOVO DISCOVERY OF MUTATED DRIVER PATHWAYS IN CANCER[1]

BY BINGHUI LIU[*,†], CHONG WU[†,2], XIAOTONG SHEN[†] AND WEI PAN[†]

Northeast Normal University[*] and University of Minnesota[†]

Next-generation sequencing studies on cancer somatic mutations have discovered that driver mutations tend to appear in most tumor samples, but they barely overlap in any single tumor sample, presumably because a single driver mutation can perturb the whole pathway. Based on the corresponding new concepts of coverage and mutual exclusivity, new methods can be designed for de novo discovery of mutated driver pathways in cancer. Since the computational problem is a combinatorial optimization with an objective function involving a discontinuous indicator function in high dimension, many existing optimization algorithms, such as a brute force enumeration, gradient descent and Newton's methods, are practically infeasible or directly inapplicable. We develop a new algorithm based on a novel formulation of the problem as nonconvex programming and nonconvex regularization. The method is computationally more efficient, effective and scalable than existing Monte Carlo searching and several other algorithms, which have been applied to The Cancer Genome Atlas (TCGA) project. We also extend the new method for integrative analysis of both mutation and gene expression data. We demonstrate the promising performance of the new methods with applications to three cancer datasets to discover de novo mutated driver pathways.

REFERENCES

BEROUKHIM, R., GETZ, G., NGHIEMPHU, L., BARRETINA, J., HSUEH, T., LINHART, D., VIVANCO, I., LEE, J. C., HUANG, J. H., ALEXANDER, S., DU, J., KAU, T., THOMAS, R. K., SHAH, K., SOTO, H., PERNER, S., PRENSNER, J., DEBIASI, R. M., DEMICHELIS, F., HATTON, C., RUBIN, M. A., GARRAWAY, L. A., NELSON, S. F., LIAU, L., MISCHEL, P. S., CLOUGHESY, T. F., MEYERSON, M., GOLUB, T. A., LANDER, E. S., MELLINGHOFF, I. K. and SELLERS, W. R. (2007). Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc. Natl. Acad. Sci. USA* **104** 20007–20012.

BOCA, S. M., KINZLER, K. W., VELCULESCU, V. E., VOGELSTEIN, B. and PARMIGIANI, G. (2010). Patient-oriented gene set analysis for cancer mutation data. *Genome Biol.* **11** R112.

BRENNAN, C. W., VERHAAK, R. G., MCKENNA, A. et al. (2013). The somatic genomic landscape of glioblastoma. *Cell* **155** 462–477.

CIRIELLO, G., CERAMI, E., SANDER, C. and SCHULTZ, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22** 398–406.

DA CUNHA SANTOS, G., SHEPHERD, F. A. and TSAO, M. S. (2011). EGFR mutations and lung cancer. *Annu. Rev. Phytopathol.* **6** 49–69.

Key words and phrases. DNA sequencing, driver mutations, optimization, subset selection, truncated L_1 penalty.

DING, L., GETZ, G., WHEELER, D. A. et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455** 1069–1075.

EFRONI, S. (2011). Detecting cancer gene networks characterized by recurrent genomic alterations in a population. *PLoS ONE* **6** e14437.

FENG, J., KIM, S. T., LIU, W. et al. (2012). An integrated analysis of germline and somatic, genetic and epigenetic alterations at 9p21.3 in glioblastoma. *Cancer* **118** 232–240.

FORBES, S. A., BEARE, D., GUNASEKARAN, P. et al. (2015). COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43** D805–D811.

FRATTINI, V., TRIFONOV, V., CHAN, J. M. et al. (2013). The integrated landscape of driver genomic alterations in glioblastoma. *Nat. Genet.* **45** 1141–1149.

GETZ, G., HÖFLING, H., MESIROV, J. P., GOLUB, T. R., MEYERSON, M., TIBSHIRANI, R. and LANDER, E. S. (2007). Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science* **317** 1500.

GILL, R. K., YANG, S. H., MEERZAMAN, D. et al. (2011). Frequent homozygous deletion of the LKB1/STK11 gene in non-small cell lung cancer. *Oncogene* **30** 3784–3791.

HARTMANN, C., BARTELS, G., GEHLHAAR, C., HOLTKAMP, N. and VON DEIMLING, A. (2005). PIK3CA mutations in glioblastoma multiforme. *Acta Neuropathol.* **109** 639–642.

HEINEMANN, V., STINTZING, S., KIRCHNER, T., BOECK, S. and JUNG, A. (2009). Clinical relevance of EGFR- and KRAS-status in colorectal cancer patients treated with monoclonal antibodies directed against the EGFR. *Cancer Treat. Rev.* **35** 262–271.

JONES, S., ZHANG, X., PARSONS, D. W. et al. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321** 1801–1806.

KANDOTH, C., MCLELLAN, M. D., VANDIN, F. et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature* **502** 333–339.

LEISERSON, M. DM., BLOKH, D., SHARAN, R. and RAPHAEL, B. J. (2013). Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.* **9** e1003054.

LI, C. and LI, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24** 1175–1182.

LIU, L., LEI, J., WILLSEY, A. et al. (2014). DAWN: A framework to identify autism genes and subnetworks using gene expression and genetics. *Molecular Autism* **5** 22.

LO, Y. L., HSIAO, C. F., JOU, Y. S. et al. (2008). ATM polymorphisms and risk of lung cancer among never smokers. *Lung Cancer* **69** 148–154.

MARDIS, E. R. and WILSON, R. K. (2009). Cancer genome sequencing: A review. *Hum. Mol. Genet.* **18** R163–R168.

MASICA, D. L. and KARCHIN, R. (2011). Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res.* **71** 4550–4561.

MEYERSON, M., GABRIEL, S. and GETZ, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11** 685–696.

MILLER, C. A., SETTLE, S. H., SULMAN, E. P., ALDAPE, K. D. and MILOSAVLJEVIC, A. (2011). Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med Genomics* **4** 34.

QIU, Y.-Q., ZHANG, S., ZHANG, X.-S. and CHEN, L. (2010). Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC Bioinform.* **11** 26.

SCHAID, D. J., SINNWELL, J. P., JENKINS, G. D., MCDONNELL, S. K., INGLE, J. N., KUBO, M., GOSS, P. E., COSTANTINO, J. P., WICKERHAM, D. L. and WEINSHILBOUM, R. M. (2012). Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. *Genet. Epidemiol.* **36** 3–16.

SCHWARTZENTRUBER, J., KORSHUNOV, A., LIU, X. Y. et al. (2012). Driver mutations in histone H3.3 and chromatin remodeling genes in paediatric glioblastoma. *Nature* **482** 226–231.

SHEN, X., PAN, W. and ZHU, Y. (2012). Likelihood-based selection and sharp parameter estimation. *J. Amer. Statist. Assoc.* **107** 223–232. MR2949354

SHOR, N. Z. (1985). *Minimization Methods for Nondifferentiable Functions*. *Springer Series in Computational Mathematics* **3**. Springer, Berlin. MR0775136

STARK, A. M., WITZEL, P., STREGE, R. J., HUGO, H. H. and MEHDORN, H. M. (2003). P53, mdm2, EGFR, and msh2 expression in paired initial and recurrent glioblastoma multiforme. *J. Neurol. Neurosurg. Psychiatry* **74** 779–783.

STURM, D., BENDER, S., JONES, D. T. W., LICHTER, P., GRILL, J., BECHER, O., HAWKINS, C., MAJEWSKI, J., JONES, C., COSTELLO, J. F., IAVARONE, A., ALDAPE, K., BRENNAN, C. W., JABADO, N. and PFISTER, S. M. (2014). Paediatric and adult glioblastoma: Multiform (epi)genomic culprits emerge. *Nat. Rev. Cancer* **14** 92–107.

THE CANCER GENOME ATLAS RESEARCH NETWORK (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455** 1061–1068.

TORKAMANI, A., TOPO, E. J. and SCHORK, N. J. (2008). Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* **92** 265–272.

TURCAN, S., ROHLE, D., GOENKA, A. et al. (2012). IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* **483** 479–483.

VANDIN, F., UPFAL, E. and RAPHAEL, B. J. (2012). De novo discovery of mutated driver pathways in cancer. *Genome Res.* **22** 375–385.

VOGELSTEIN, B. and KINZLER, K. W. (2004). Cancer genes and the pathways they control. *Nat. Med.* **10** 789–799.

WANG, K., LI, M. and BUCAN, M. (2007). Pathway-based approaches for analysis of genome-wide association studies. *Am. J. Hum. Genet.* **81** 1278–1283.

ZHANG, S. and ZHOU, X. J. (2014). Matrix factorization methods for integrative cancer genomics. *Methods Mol. Biol.* **1176** 229–242.

ZHAO, J., ZHANG, S., WU, L. and ZHANG, X. (2012). Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics* **28** 2940–2947.

ZHUANG, G., SONG, W., AMATO, K. et al. (2012). Effects of cancer-associated EPHA3 mutations on lung cancer. *J. Natl. Cancer Inst.* **104** 1182–1197.

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1513–1536
DOI: 10.1214/17-AOAS1044

LATENT CLASS MODELING USING MATRIX COVARIATES WITH APPLICATION TO IDENTIFYING EARLY PLACEBO RESPONDERS BASED ON EEG SIGNALS[1]

BY BEI JIANG*, EVA PETKOVA[†,‡], THADDEUS TARPEY[§] AND R. TODD OGDEN[¶]

University of Alberta, New York University[†], Nathan S. Kline Institute for Psychiatric Research[‡], Wright State University[§] and Columbia University[¶]*

Latent class models are widely used to identify unobserved subgroups (i.e., latent classes) based upon one or more manifest variables. The probability of belonging to each subgroup is typically modeled as a function of a set of measured covariates. In this paper, we extend existing latent class models to incorporate matrix covariates. This research is motivated by a randomized placebo-controlled depression clinical trial. One study goal is to identify a subgroup of subjects who experience symptoms improvement early on during antidepressant treatment, which is considered to be an indication of a placebo rather than a true pharmacological response. We want to relate the likelihood of belonging to this subgroup of early responders to baseline electroencephalography (EEG) measurement that takes the form of a matrix. The proposed method is built upon a low-rank Candecomp/Parafac (CP) decomposition of the target coefficient matrix through low-dimensional latent variables, which effectively reduces the model dimensionality. We adopt a Bayesian hierarchical modeling approach to estimate the latent variables, which allows a flexible way to incorporate prior knowledge about covariate effect heterogeneity and offers a data-driven method of regularization. Simulation studies suggest that the proposed method is robust against potentially misspecified rank in the CP decomposition. With the motivating example we show how the proposed method can be applied to extract valuable information from baseline EEG measurements that explains the likelihood of belonging to the early responder subgroup, helping to identify placebo responders and suggesting new targets for the study of placebo response.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19** 716–723. MR0423716

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

BANDEEN-ROCHE, K., MIGLIORETTI, D. L., ZEGER, S. L. and RATHOUZ, P. J. (1997). Latent variable regression for multiple discrete outcomes. *J. Amer. Statist. Assoc.* **92** 1375–1386. MR1615248

BESAG, J. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746. MR1380811

Key words and phrases. Candecomp/Parafac (CP) matrix decomposition, Bayesian hierarchical modeling, data-driven regularization, major depression, placebo effect.

BONATE, P. L. and HOWARD, D. R. (2011). *Pharmacokinetics in Drug Development. Advances and Applications* **3**. Springer, Berlin.

BRUDER, G. E., STEWART, J. W., TENKE, C. E., MCGRATH, P. J., LEITE, P., BHATTACHARYA, N. and QUITKIN, F. M. (2001). Electroencephalographic and perceptual asymmetry differences between responders and nonresponders to an SSRI antidepressant. *Biol. Psychiatry* **49** 416–425.

BRUDER, G. E., SEDORUK J. P., STEWART J. W., MCGRATH, P., QUITKIN, F. M. and TENKE, C. E. (2008). EEG alpha measures predict therapeutic response to an SSRI antidepressant: Pre and post treatment findings. *Biol. Psychiatry* **63** 1171.

CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90** 1313–1321.

CIARLEGLIO, A., PETKOVA, E., OGDEN, R. T. and TARPEY, T. (2015). Treatment decisions based on scalar and functional baseline covariates. *Biometrics* **71** 884–894. MR3436714

CLOGG, C. C. (1995). Latent class models. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (G. Arminger, C. C. Clogg and M. E. Sobel, eds.) 311–359. Plenum Press, New York.

COLLINS, L. M. and LANZA, S. T. (2013). *Latent Class and Latent Transition Analysis*: *With Applications in The Social*, *Behavioral*, *and Health Sciences*. Wiley, Hoboken, NJ.

ELLIOTT, M. R. (2007). Identifying latent clusters of variability in longitudinal data. *Biostatistics* **8** 756–771.

ELLIOTT, M. R., GALLO, J. J., TEN HAVE, T. R., BOGNER, H. R. and KATZ, I. R. (2005). Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction. *Biostatistics* **6** 119–143.

GARRETT, E. S. and ZEGER, S. L. (2000). Latent class model diagnosis. *Biometrics* **56** 1055–1067. MR1815583

GEISSER, S. and EDDY, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74** 153–160. MR0529531

GELFAND, A. E. and DEY, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **56** 501–514. MR1278223

GELMAN, A., HWANG, J. and VEHTARI, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24** 997–1016. MR3253850

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GORMLEY, I. C. and MURPHY, T. B. (2011). Mixture of experts modelling with social science applications. In *Mixtures*: *Estimation and Applications*. *Wiley Ser. Probab. Stat.* 101–121. Wiley, Chichester. MR2883359

HOLSBOER, F. (2008). How can we realize the promise of personalized antidepressant medicines? *Nat. Rev., Neurosci.* **9** 638–646.

HUNG, H. and WANG, C.-C. (2013). Matrix variate logistic regression model with application to EEG data. *Biostatistics* **14** 189–202.

JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. and HINTON, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput.* **3** 79–87.

JIANG, B., ELLIOTT, M. R., SAMMEL, M. D. and WANG, N. (2015). Joint modeling of cross-sectional health outcomes and longitudinal predictors via mixtures of means and variances. *Biometrics* **71** 487–497. MR3366253

JIANG, B., PETKOVA, E., TARPEY, T. and OGDEN, R. T (2017). Supplement to "Latent class modeling using matrix covariates with application to identifying early placebo responders based on EEG signals." DOI:10.1214/17-AOAS1044SUPP.

JORDAN, M. I. and JACOBS, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Comput.* **6** 181–214.

JOYCE, P. R. and PAYKEL, E. S. (1989). Predictors of drug response in depression. *Arch. Gen. Psychiatry* **46** 89–99.

KAMARAJAN, C., PANDEY, A. K., CHORLIAN, D. B. and PORJESZ, B. (2015). The use of current source density as electrophysiological correlates in neuropsychiatric disorders: A review of human studies. *Int. J. Psychophysiol.* **97** 310–322.

KHODAYARI-ROSTAMABAD, A., REILLY, J. P., HASEY, G., DEBRUIN, H. and MACCRIMMON, D., (2010). Using pre-treatment EEG data to predict response to SSRI treatment for MDD. In *Proceedings of the* 2010 32*nd Annual International Conference of the IEEE Engineering in Medicine and Biology* (*EBMC* 2010) 6103–6106. IEEE, New York. DOI:10.1109/IEMBS.2010.5627823.

KIM, S., CHEN, M.-H. and DEY, D. K. (2008). Flexible generalized t-link models for binary response data. *Biometrika* **95** 93–106. MR2409717

KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056

LAZARSFELD, P. F. and HENRY, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston, MA.

LEUCHTER, A. F., COOK, I. A., WITTE, E. A., MORGAN, M. and ABRAMS, M. (2002). Changes in brain function of depressed subjects during treatment with placebo. *Am. J. Psychiatr.* **159** 122–129.

LI, B., KIM, M. K. and ALTMAN, N. (2010). On dimension folding of matrix- or array-valued statistical objects. *Ann. Statist.* **38** 1094–1121. MR2604706

LU, H., PLATANIOTIS, K. N. and VENETSANOPOULOS, A. N. (2008). MPCA: Multilinear principal component analysis of tensor objects. *IEEE Trans. Neural Netw.* **19** 18–39.

MACCUTCHEON, A. L. (1987). *Latent Class Analysis*. Sage Publications, Thousand Oaks, CA.

MATTHEWS, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **405** 442–451.

MUMTAZ, W., MALIK, A. S., YASIN, M. A. M. and XIA, L. (2015). Review on EEG and ERP predictive biomarkers for major depressive disorder. *Biomed. Signal Process. Control* **22** 85–98.

MUTHÉN, B. and BROWN, H. C. (2009). Estimating drug effects in the presence of placebo response: Causal inference using growth mixture modeling. *Stat. Med.* **28** 3363–3385. MR2744369

MUTHÉN, B. and SHEDDEN, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55** 463–469.

NEELON, B., O'MALLEY, A. J. and NORMAND, S.-L. T. (2011). A Bayesian two-part latent class model for longitudinal medical expenditure data: Assessing the impact of mental health and substance abuse parity. *Biometrics* **67** 280–289. MR2898840

NUNEZ, P. L. and SRINIVASAN, R. (2006). *Electric Fields of the Brain*: *The Neurophysics of EEG*. Oxford Univ. Press, New York.

PATEL, M. J., KHALAF, A. and AIZENSTEIN, H. J. (2016). Studying depression using imaging and machine learning methods. *NeuroImage*: *Clinical* **10** 115–123.

PETKOVA, E., TARPEY, T. and GOVINDARAJULU, U. (2009). Predicting potential placebo effect in drug treated subjects. *Int. J. Biostat.* **5** Art. 23, 27. MR2533809

PHILLIPS, M. L., CHASE, H. W., SHELINE, Y. I., ETKIN, T ALMEIDA J. R, A., DECKERSBACH, T. and TRIVEDI, M. H. (2015). Identifying predictors, moderators, and mediators of antidepressant response in major depressive disorder: Neuroimaging approaches. *Am. J. Psychiatr.* **172** 124–138.

POWERS, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.* **2** 37–63.

QUITKIN, F. M., MCGRATH, P. J., RABKIN, J. G., STEWART, J. W., HARRISON, W., ROSS, D. C., TRICAMO, E., FLEISS, J., MARKOWITZ, J. and KLEIN, D. F. (1991). Different types of placebo response in patients receiving antidepressants. *Am. J. Psychiatr.* **148** 197–203.

SHEN, J. and HE, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *J. Amer. Statist. Assoc.* **110** 303–312. MR3338504

SING, T., SANDER, O. BEERENWINKEL, N.and LENGAUER, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* **21** 3940–3941.

SONAWALLA, S. B. and ROSENBAUM, J. F. (2002). Placebo response in depression. *Dialogues Clin. Neurosci.* **4** 105–113.

STEWART, J. W., QUITKIN, F. M., MCGRATH, P. J., AMSTERDAM, J., FAVA, M., FAWCETT, J., REIMHERR, F., ROSENBAUM, J., BEASLEY, C. and ROBACK, P. (1998). Use of pattern analysis to predict differential relapse of remitted patients with major depression during 1 year of treatment with fluoxetine or placebo. *Arch. Gen. Psychiatry* **55** 334–343.

TARPEY, T. and PETKOVA, E. (2010). Latent regression analysis. *Stat. Model.* **10** 133–158. MR2649773

TARPEY, T., PETKOVA, E. and OGDEN, R. T. (2003). Profiling placebo responders by self-consistent partitioning of functional data. *J. Amer. Statist. Assoc.* **98** 850–858. MR2055493

TARPEY, T., YUN, D. and PETKOVA, E. (2008). Model misspecification: Finite mixture or homogeneous? *Stat. Model.* **8** 199–218. MR2750637

TENKE, C. E., KAYSER, J., MANNA, C. G., FEKRI, S., KROPPMANN, C. J., SCHALLER, J. D., ALSCHULER, D. M., STEWART, J. W., MCGRATH, P. J. and BRUDER, G. E. (2011). Current source density measures of electroencephalographic alpha predict antidepressant treatment response. *Biol. Psychiatry* **70** 388–394.

VEHTARI, A. and OJANEN, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Stat. Surv.* **6** 142–228. MR3011074

WADE, E. C. and IOSIFESCU, D. V. (2016). Using EEG for treatment guidance in major depressive disorder. *Biol. Psychiatry*: *Cognitive Neurosci. Neuroimag.* **1** 411–422.

WATANABE, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11** 3571–3594. MR2756194

WHITE, A. and MURPHY, T. B. (2016). Mixed-membership of experts stochastic blockmodel. *Netw. Sci.* **4** 48–80.

ZHOU, H., LI, L. and ZHU, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Amer. Statist. Assoc.* **108** 540–552. MR3174640

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1537–1560
DOI: 10.1214/17-AOAS1045

A MULTI-STATE CONDITIONAL LOGISTIC REGRESSION MODEL FOR THE ANALYSIS OF ANIMAL MOVEMENT[1]

BY AURÉLIEN NICOSIA, THIERRY DUCHESNE, LOUIS-PAUL RIVEST AND DANIEL FORTIN

Université Laval

A multi-state version of an animal movement analysis method based on conditional logistic regression, called Step Selection Function (SSF), is proposed. In ecology SSF is developed from a comparison between the observed location of an animal and randomly sampled locations at each time step. Interpretation of the parameters in the multi-state model and the impact of different sampling schemes for the random locations are discussed. We prove the relationship between the new model, called HMM-SSF, and a random walk model on the plane. This relationship allows one to use both movement characteristics and local discrete choice behaviors when identifying the model's hidden states. The new HMM-SSF is used to model the movement behavior of GPS-collared bison in Prince Albert National Park, Canada, where it successfully teases apart areas used to forage and to travel. The analysis thus provides valuable insights into how bison adjust their movement to habitat features, thereby revealing spatial determinants of functional connectivity in heterogeneous landscapes.

REFERENCES

ALBERTSEN, C. M., WHORISKEY, K., YURKOWSKI, D., NIELSEN, A. and FLEMMING, J. M. (2015). Fast fitting of non-Gaussian state–space models to animal movement data via template model builder. *Ecology* **96** 2598–2604.

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.

AVRIEL, M. (2003). *Nonlinear Programming*: *Analysis and Methods*. Dover Publications, Inc., Mineola, NY. MR2015090

BASILLE, M., FORTIN, D., DUSSAULT, C., OUELLET, J.-P. and COURTOIS, R. (2012). Ecologically based definition of seasons clarifies predator–prey interactions. *Ecography* **36** 220–229.

BASILLE, M., FORTIN, D., DUSSAULT, C., BASTILLE-ROUSSEAU, G., OUELLET, J.-P. and COURTOIS, R. (2015). Plastic response of fearful prey to the spatiotemporal dynamics of predator distribution. *Ecology* **96** 2622–2631.

BAUM, L. E. and PETRIE, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* **37** 1554–1563. MR0202264

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

Key words and phrases. Animal movement, biased correlated random walk, conditional logistic regression, GPS, hidden Markov model, Step Selection Function.

COURBIN, N., FORTIN, D., DUSSAULT, C. and COURTOIS, R. (2014). Logging-induced changes in habitat network connectivity shape behavioral interactions in the wolf–caribou–moose system. *Ecol. Monogr.* **84** 265–285.

DANCOSE, K., FORTIN, D. and GUO, X. (2011). Mechanisms of functional connectivity: The case of free-ranging bison in a forest landscape. *Ecol. Appl.* **21** 1871–1885.

DUCHESNE, T., FORTIN, D. and COURBIN, N. (2010). Mixed conditional logistic regression for habitat selection studies. *J. Anim. Ecol.* **79** 548–555.

DUCHESNE, T., FORTIN, D. and RIVEST, L.-P. (2015). Equivalence between step selection functions and biased correlated random walks for statistical inference on animal movement. *PLoS ONE* **10** e0122947.

FORESTER, J. D., IM, H. K. and RATHOUZ, P. J. (2009). Accounting for animal movement in estimation of resource selection functions: Sampling and data analysis. *Ecology* **90** 3554–3565.

FORTIN, M. E. (2007). Effets de la taille du groupe sur la sélection de l'habitat à plusieurs échelles spatio-temporelles par le bison des plaines (Bison bison bison). M.Sc. thesis, Université Laval, Quebec, Canada.

FORTIN, D., BEYER, H. L., BOYCE, M. S., SMITH, D. W., DUCHESNE, T. and MAO, J. S. (2005). Wolves influence elk movements: Behavior shapes a trophic cascade in Yellowstone National Park. *Ecology* **86** 1320–1330.

FRÜHWIRTH-SCHNATTER, S. (2013). *Finite Mixture and Markov Switching Models*. Springer, New York.

FRYXELL, J. M., HAZELL, M., BORGER, L., DALZIEL, B. D., HAYDON, D. T., MORALES, J. M., MCINTOSH, T. and ROSATTE, R. C. (2008). Multiple movement modes by large herbivores at multiple spatiotemporal scales. *Proc. Natl. Acad. Sci. USA* **105** 19114–19119.

HOLYOAK, M., CASAGRANDI, R., NATHAN, R., REVILLA, E. and SPIEGEL, O. (2008). Trends and missing parts in the study of movement ecology. *Proc. Natl. Acad. Sci. USA* **105** 19060–19065.

HOLZMANN, H., MUNK, A., SUSTER, M. and ZUCCHINI, W. (2006). Hidden Markov models for circular and linear–circular time series. *Environ. Ecol. Stat.* **13** 325–347. MR2242193

HOSMER, D. W. and LEMESHOW, S. (2000). *Applied Logistic Regression*. Wiley, New York.

JONSEN, I. D., FLEMMING, J. M. and MYERS, R. A. (2005). Robuts state–space modeling of animal movement data. *Ecology* **86** 2874–2880.

KINDLMANN, P. and BUREL, F. (2008). Connectivity measures: A review. *Landsc. Ecol.* **23** 879–890.

LANGROCK, R., KING, R., MATTHIOPOULOS, J., THOMAS, L., FORTIN, D. and MORALES, J. M. (2012). Flexible and practical modeling of animal telemetry data: Hidden Markov models and extensions. *Ecology* **93** 2336–2342.

LATOMBE, G., PARROTT, L., BASILLE, M. and FORTIN, D. (2014). Uniting statistical and individual-based approaches for animal movement modelling. *PLoS ONE* **9** e99938.

LEHMANN, E. L. and CASELLA, G. (2003). *Theory of Point Estimation*, 2nd ed. Springer, New York.

MARDIA, K. V. and JUPP, P. E. (1999). *Directional Statistics*. Wiley, Chichester.

MORALES, J. M., HAYDON, D. T., FRAIR, J., HOLSINGER, K. E. and FRYXELL, J. M. (2004). Extracting more out of relocation data: Building movement models as mixtures of random walks. *Ecology* **85** 2436–2445.

MURTAUGH, P. A. (2007). Simplicity and complexity in ecological data analysis. *Ecology* **88** 56–62.

NATHAN, R., GETZ, W. M., REVILLA, E., HOLYOAK, M., KADMON, R., SALTZ, D. and SMOUSE, P. E. (2008). A movement ecology paradigm for unifying organismal movement research. *Proc. Natl. Acad. Sci. USA* **105** 19052–19059.

NICOSIA, A., DUCHESNE, T., RIVEST, L.-P. and FORTIN, D. (2017a). A general hidden state random walk model for animal movement. *Comput. Statist. Data Anal.* **105** 76–95. MR3552190

NICOSIA, A., DUCHESNE, T., RIVEST, L.-P. and FORTIN, D. (2017b). Supplement to "A multi-state conditional logistic regression model for the analysis of animal movement." DOI:10.1214/17-AOAS1045SUPPA.

NICOSIA, A., DUCHESNE, T., RIVEST, L.-P. and FORTIN, D. (2017c). Supplement to "A multi-state conditional logistic regression model for the analysis of animal movement." DOI:10.1214/17-AOAS1045SUPPB.

NICOSIA, A., DUCHESNE, T., RIVEST, L.-P. and FORTIN, D. (2017d). Supplement to "A multi-state conditional logistic regression model for the analysis of animal movement." DOI:10.1214/17-AOAS1045SUPPC.

NICOSIA, A., DUCHESNE, T., RIVEST, L.-P. and FORTIN, D. (2017e). Supplement to "A multi-state conditional logistic regression model for the analysis of animal movement." DOI:10.1214/17-AOAS1045SUPPD.

PATIL, G. P. (2005). Weighted Distributions. In *Encyclopedia of Biostatistics*. Wiley, New York.

PATTERSON, T., THOMAS, L., WILCOX, C., OVASKAINEN, O. and MATTHIOPOULOS, J. (2008). State–space models of individual animal movement. *Trends Ecol. Evol.* **23** 87–94.

PROKOPENKO, C. M., BOYCE, M. S. and AVGAR, T. (2016). Characterizing wildlife behavioural responses to roads using integrated step selection analysis. *J. Appl. Ecol.* **54** 470–479.

RIVEST, L.-P., DUCHESNE, T., NICOSIA, A. and FORTIN, D. (2016). A general angular regression model for the analysis of data on animal movement in ecology. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 445–463. MR3470586

THERNEAU, T. M. (2015). A Package for Survival Analysis in S, version 2.38.

TISCHENDORF, L. and FAHRIG, L. (2000). On the usage and measurement of landscape connectivity. *Oikos* **90** 7–19.

TRAIN, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge Univ. Press, Cambridge. MR2003007

TURCHIN, P. (1998). *Quantitative Analysis of Movement*: *Measuring and Modeling Population Redistribution in Animals and Plants*. Sinauer Associates, Sunderland, MA, USA.

VANAK, A. T., FORTIN, D., THAKER, M., OGDEN, M., OWEN, C., GREATWOOD, S. and SLOTOW, R. (2013). Moving to stay in place: Behavioral mechanisms for coexistence of African large carnivores. *Ecology* **94** 2619–2631.

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1561–1592
DOI: 10.1214/17-AOAS1046

BAYESIAN LARGE-SCALE MULTIPLE REGRESSION WITH SUMMARY STATISTICS FROM GENOME-WIDE ASSOCIATION STUDIES[1]

BY XIANG ZHU AND MATTHEW STEPHENS

University of Chicago

Bayesian methods for large-scale multiple regression provide attractive approaches to the analysis of genome-wide association studies (GWAS). For example, they can estimate heritability of complex traits, allowing for both polygenic and sparse models; and by incorporating external genomic data into the priors, they can increase power and yield new biological insights. However, these methods require access to individual genotypes and phenotypes, which are often not easily available. Here we provide a framework for performing these analyses without individual-level data. Specifically, we introduce a "Regression with Summary Statistics" (RSS) likelihood, which relates the multiple regression coefficients to univariate regression results that are often easily available. The RSS likelihood requires estimates of correlations among covariates (SNPs), which also can be obtained from public databases. We perform Bayesian multiple regression analysis by combining the RSS likelihood with previously proposed prior distributions, sampling posteriors by Markov chain Monte Carlo. In a wide range of simulations RSS performs similarly to analyses using the individual data, both for estimating heritability and detecting associations. We apply RSS to a GWAS of human height that contains 253,288 individuals typed at 1.06 million SNPs, for which analyses of individual-level data are practically impossible. Estimates of heritability (52%) are consistent with, but more precise, than previous results using subsets of these data. We also identify many previously unreported loci that show evidence for association with height in our analyses. Software is available at https://github.com/stephenslab/rss.

REFERENCES

1000 GENOMES PROJECT CONSORTIUM (2010). A map of human genome variation from population-scale sequencing. *Nature* **467** 1061–1073.

AQEILAN, R. I., HASSAN, M. Q., DE BRUIN, A., HAGAN, J. P., VOLINIA, S., PALUMBO, T., HUSSAIN, S., LEE, S.-H., GAUR, T., STEIN, G. S. et al. (2008). The *WWOX* tumor suppressor is essential for postnatal survival and normal bone metabolism. *J. Biol. Chem.* **283** 21629–21639.

BOOS, D. D. (1985). A converse to Scheffé's theorem. *Ann. Statist.* **13** 423–427. MR0773179

BULIK-SULLIVAN, B., LOH, P.-R., FINUCANE, H., RIPKE, S., YANG, J., PSYCHIATRIC GENOMICS CONSORTIUM SCHIZOPHRENIA WORKING GROUP, PATTERSON, N., DALY, M. J., PRICE, A. L. and NEALE, B. M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47** 291–295.

Key words and phrases. Summary statistics, Bayesian regression, genome wide, association study, multiple-SNP analysis, variable selection, heritability, explained variation, Markov chain Monte Carlo.

CARBONETTO, P. and STEPHENS, M. (2013). Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in crohn's disease. *PLoS Genet*. **9** e1003770.

CASELLA, G. and ROBERT, C. P. (1996). Rao–Blackwellisation of sampling schemes. *Biometrika* **83** 81–94. MR1399157

CHEN, W., LARRABEE, B. R., OVSYANNIKOVA, I. G., KENNEDY, R. B., HARALAMBIEVA, I. H., POLAND, G. A. and SCHAID, D. J. (2015). Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* **200** 719–736.

DEL MARE, S., KUREK, K. C., STEIN, G. S., LIAN, J. B. and AQEILAN, R. I. (2011). Role of the *WWOX* tumor suppressor gene in bone homeostasis and the pathogenesis of osteosarcoma. *Am. J. Cancer Res*. **1** 585.

DEVLIN, B. and ROEDER, K. (1999). Genomic control for association studies. *Biometrics* **55** 997–1004.

DONNELLY, P. (2008). Progress and challenges in genome-wide association studies in humans. *Nature* **456** 728–731.

EFRON, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80** 3–26. MR1225211

EHRET, G. B., LAMPARTER, D., HOGGART, C. J., WHITTAKER, J. C., BECKMANN, J. S., KUTALIK, Z., GENETIC INVESTIGATION OF ANTHROPOMETRIC TRAITS CONSORTIUM et al. (2012). A multi-SNP locus-association method reveals a substantial fraction of the missing heritability. *Am. J. Hum. Genet.* **91** 863–871.

EVANGELOU, E. and IOANNIDIS, J. P. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet*. **14** 379–389.

FINUCANE, H. K., BULIK-SULLIVAN, B., GUSEV, A., TRYNKA, G., RESHEF, Y., LOH, P.-R., ANTTILA, V., XU, H., ZANG, C., FARH, K. et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet*. **47** 1228–1235.

FRAZER, K. A., BALLINGER, D. G., COX, D. R., HINDS, D. A., STUVE, L. L., GIBBS, R. A., BELMONT, J. W., BOUDREAU, A., HARDENBOL, P., LEAL, S. M. et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449** 851–861.

GLOBAL LIPIDS GENETICS CONSORTIUM (2013). Discovery and refinement of loci associated with lipids levels. *Nat. Genet.* **45** 1274–1283.

GUAN, Y. and STEPHENS, M. (2008). Practical issues in imputation-based association mapping. *PLoS Genet*. **4** e1000279.

GUAN, Y. and STEPHENS, M. (2011). Bayesian variable selection regression for Genome-wide association studies and other large-scale problems. *Ann. Appl. Stat*. **5** 1780–1815. MR2884922

GUAN, Y. and WANG, K. (2013). Whole-genome multi-SNP-phenotype association analysis. In *Advances in Statistical Bioinformatics* 224–243. Cambridge Univ. Press, Cambridge. MR3155921

GUSEV, A., LEE, S. H., TRYNKA, G., FINUCANE, H., VILHJÁLMSSON, B. J., XU, H., ZANG, C., RIPKE, S., BULIK-SULLIVAN, B., STAHL, E., KÄHLER, A. K., HULTMAN, C. M., PURCELL, S. M., MCCARROLL, S. A., DALY, M., PASANIUC, B., SULLIVAN, P. F., NEALE, B. M., WRAY, N. R., RAYCHAUDHURI, S. and PRICE, A. L. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet*. **95** 535–552.

HOGGART, C. J., WHITTAKER, J. C., IORIO, M. D. and BALDING, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet*. **4** e1000130.

HORMOZDIARI, F., KOSTEM, E., KANG, E. Y., PASANIUC, B. and ESKIN, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics* **198** 497–508.

IIOKA, T., FURUKAWA, K., YAMAGUCHI, A., SHINDO, H., YAMASHITA, S. and TSUKAZAKI, T. (2003). P300/CBP acts as a coactivator to cartilage homeoprotein-1 (Cart1), paired-like homeoprotein, through acetylation of the conserved lysine residue adjacent to the homeodomain. *J. Bone Miner. Res*. **18** 1419–1429.

KANG, H. M., SUL, J. H., SERVICE, S. K., ZAITLEN, N. A., KONG, S.-Y., FREIMER, N. B., SABATTI, C., ESKIN, E. et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42** 348–354.

KURKÓ, J., BESENYEI, T., LAKI, J., GLANT, T. T., MIKECZ, K. and SZEKANECZ, Z. (2013). Genetics of rheumatoid arthritis—A comprehensive review. *Clin. Rev. Allergy Immunol.* **45** 170–179.

LEE, D., BIGDELI, T. B., RILEY, B. P., FANOUS, A. H. and BACANU, S.-A. (2013). DIST: Direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics* **29** 2925–2927.

LEE, D., WILLIAMSON, V. S., BIGDELI, T. B., RILEY, B. P., FANOUS, A. H., VLADIMIROV, V. I. and BACANU, S.-A. (2015). JEPEG: A summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics* **31** 1176–1182.

LI, D., SAKUMA, R., VAKILI, N. A., MO, R., PUVIINDRAN, V., DEIMLING, S., ZHANG, X., HOPYAN, S. and HUI, C.-C. (2014). Formation of proximal and anterior limb skeleton requires early function of *I*rx3 and *I*rx5 and is negatively regulated by Shh signaling. *Dev. Cell* **29** 233–240.

LIAO, W.-J., TSAO, K.-C. and YANG, R.-B. (2016). Electrostatics and N-glycan-mediated membrane tethering of *SCUBE1* is critical for promoting bone morphogenetic protein signalling. *Biochem. J.* **473** 661–672.

LIN, D. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* **21** 781–787.

LIU, J. Z., MCRAE, A. F., NYHOLT, D. R., MEDLAND, S. E., WRAY, N. R., BROWN, K. M., HAYWARD, N. K., MONTGOMERY, G. W., VISSCHER, P. M., MARTIN, N. G. et al. (2010). A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87** 139–145.

LOH, P.-R., TUCKER, G., BULIK-SULLIVAN, B. K., VILHJALMSSON, B. J., FINUCANE, H. K., CHASMAN, D. I., RIDKER, P. M., NEALE, B. M., BERGER, B., PATTERSON, N. et al. (2015). Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nat. Genet.* **47** 284–290.

MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. and DONNELLY, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39** 906–913.

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. and HIRSCHHORN, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9** 356–369.

MOSER, G., LEE, S. H., HAYES, B. J., GODDARD, M. E., WRAY, N. R. and VISSCHER, P. M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet.* **11** e1004969.

NATURE GENETICS (2012). Asking for more. *Nat. Genet.* **44** 733.

NEWCOMBE, J., CONTI, V. and RICHARDSON, S. (2016). JAM: a scalable bayesian framework for joint analysis of marginal SNP effects. *Genet. Epidemiol.* **40** 188–201.

PALLA, L. and DUDBRIDGE, F. (2015). A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *Am. J. Hum. Genet.* **97** 250–259.

PARK, J.-H., WACHOLDER, S., GAIL, M. H., PETERS, U., JACOBS, K. B., CHANOCK, S. J. and CHATTERJEE, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42** 570–575.

PEISE, E., FABREGAT-TRAVER, D. and BIENTINESI, P. (2015). High performance solutions for big-data GWAS. *Parallel Comput.* **42** 75–87.

PICKRELL, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94** 559–573.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.

PRICE, A. L., ZAITLEN, N. A., REICH, D. and PATTERSON, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11** 459–463.

PRITCHARD, J. K. and PRZEWORSKI, M. (2001). Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.* **69** 1–14.

SABATTI, C. (2013). Multivariate linear models for GWAS. In *Advances in Statistical Bioinformatics* (K.-A. Do, Z. S. Qin and M. Vannucci, eds.) 188–207. Cambridge Univ. Press, Cambridge. MR3155918

SEAMAN, S. R. and MÜLLER-MYHSOK, B. (2005). Rapid simulation of P values for product methods and multiple-testing adjustment in association studies. *Am. J. Hum. Genet.* **76** 399–408.

SERVIN, B. and STEPHENS, M. (2007). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genet.* **3** e114.

SMEDLEY, D., HAIDER, S., DURINCK, S., PANDINI, L., PROVERO, P., ALLEN, J., ARNAIZ, O., AWEDH, M. H., BALDOCK, R., BARBIERA, G. et al. (2015). The BioMart community portal: An innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* **43** W589–W598.

STEPHENS, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PLoS ONE* **8** e65245.

STEPHENS, M. (2017). False discovery rates: A new deal. *Biostatistics* **18** 275–294.

SWEETING, T. J. (1986). On a converse to Scheffé's theorem. *Ann. Statist.* **14** 1252–1256. MR0856821

VILHJALMSSON, B., YANG, J., FINUCANE, H. K., GUSEV, A., LINDSTROM, S., RIPKE, S., GENOVESE, G., LOH, P.-R., BHATIA, G., DO, R. et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97** 576–592.

VISSCHER, P. M., HILL, W. G. and WRAY, N. R. (2008). Heritability in the genomics era—Concepts and misconceptions. *Nat. Rev. Genet.* **9** 255–266.

WAKEFIELD, J. (2009). Bayes factors for genome-wide association studies: Comparison with P-values. *Genet. Epidemiol.* **33** 79–86.

WANG, K., LI, M. and HAKONARSON, H. (2010). Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* **11** 843–854.

WELLCOME TRUST CASE CONTROL CONSORTIUM (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* **447** 661–678.

WEN, X. and STEPHENS, M. (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Ann. Appl. Stat.* **4** 1158–1182. MR2751337

WEN, X. and STEPHENS, M. (2014). Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene-environment interactions. *Ann. Appl. Stat.* **8** 176–203. MR3191987

WOOD, A. R., ESKO, T., YANG, J., VEDANTAM, S., PERS, T. H., GUSTAFSSON, S., CHU, A. Y., ESTRADA, K., LUAN, J., KUTALIK, Z. et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46** 1173–1186.

YANG, J., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W. et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42** 565–569.

YANG, J., MANOLIO, T. A., PASQUALE, L. R., BOERWINKLE, E., CAPORASO, N., CUNNINGHAM, J. M., DE ANDRADE, M., FEENSTRA, B., FEINGOLD, E., HAYES, M. G. et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43** 519–525.

YANG, J., FERREIRA, T., MORRIS, A. P., MEDLAND, S. E., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W., WEEDON, M. N., LOOS, R. J. et al. (2012). Condi-

tional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44** 369–375.

ZHANG, H., WHEELER, W., HYLAND, P. L., YANG, Y., SHI, J., CHATTERJEE, N. and YU, K. (2016). A powerful procedure for pathway-based meta-analysis using summary statistics identifies 43 pathways associated with type II diabetes in European populations. *PLoS Genet.* **12** e1006122.

ZHOU, X., CARBONETTO, P. and STEPHENS, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* **9** e1003264.

ZHOU, X. and STEPHENS, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44** 821–824.

ZHU, X. and STEPHENS, M. (2017). Supplement to "Bayesian large-scale multiple regression with summary statistics from genome-wide association studies". DOI:10.1214/17-AOAS1046SUPP.

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1593–1616
DOI: 10.1214/17-AOAS1047

MODELING CD4$^+$ T CELLS DYNAMICS IN HIV-INFECTED PATIENTS RECEIVING REPEATED CYCLES OF EXOGENOUS INTERLEUKIN 7[1]

BY ANA JARNE[*,†,‡,§], DANIEL COMMENGES[*,†,‡,§], LAURA VILLAIN[*,†,‡,§], MÉLANIE PRAGUE[*,†,‡,§], YVES LÉVY[*,§] AND RODOLPHE THIÉBAUT[*,†,‡,§]

INSERM[*], INRIA[†], University of Bordeaux[‡] and Vaccine Research Institute[§]

Combination antiretroviral therapy successfully controls viral replication in most HIV infected patients. This is normally followed by a reconstitution of the CD4$^+$ T cells pool, but not for all patients. For these patients, an immunotherapy based on injections of Interleukin 7 (IL-7) has been recently proposed in the hope of obtaining long-term reconstitution of the T cells pool. Several questions arise as to the long-term efficiency of this treatment and the best protocol to apply. Mathematical and statistical models can help answer these questions.

We developed a model based on a system of ordinary differential equations and a statistical model of variability and measurement. We can estimate key parameters of this model using the data from the main studies for this treatment, the INSPIRE, INSPIRE 2, and INSPIRE 3 trials. In all three studies, cycles of three injections have been administered; in the last two studies, for the first time, repeated cycles of IL-7 have been administered. Repeated measures of total CD4$^+$ T cells count in 128 patients, as well as CD4$^+$Ki67$^+$ T cells count (the number of cells expressing the proliferation marker Ki67) in some of them, were available. Our aim was to estimate the possibly different effects of successive injections in a cycle, to estimate the effect of repeated cycles and to assess different protocols.

The use of dynamical models together with our complex statistical approach allow us to analyze major biological questions. We found a strong effect of IL-7 injections on the proliferation rate; however, the effect of the third injection of the cycle appears to be much weaker than the first ones. Also, despite a slightly weaker effect of repeated cycles with respect to the initial one, our simulations show the ability of this treatment of maintaining adequate CD4$^+$ T cells count for years. We also compared different protocols, showing that cycles of two injections should be sufficient in most cases.

REFERENCES

COMMENGES, D., JACQMIN-GADDA, H., PROUST, C. and GUEDJ, J. (2006). A Newton-like algorithm for likelihood maximization: The robust-variance scoring algorithm. Preprint, arXiv:math/0610402.

COMMENGES, D., JOLY, P., GÉGOUT-PETIT, A. and LIQUET, B. (2007). Choice between semiparametric estimators of Markov and non-Markov multi-state models from coarsened observations. *Scand. J. Stat.* **34** 33–52. MR2325241

Key words and phrases. Mechanistic models, Interleukin 7, HIV, modeling, CD4.

COMMENGES, D., SAYYAREH, A., LETENNEUR, L., GUEDJ, J. and BAR-HEN, A. (2008). Estimating a difference of Kullback–Leibler risks using a normalized difference of AIC. *Ann. Appl. Stat.* **2** 1123–1142. MR2522174

COMMENGES, D., PROUST-LIMA, C., SAMIERI, C. and LIQUET, B. (2015). A universal approximate cross-validation criterion for regular risk functions. *Int. J. Biostat.* **11** 51–67. MR3341512

DRYLEWICZ, J., COMMENGES, D. and THIEBAUT, R. (2012). Maximum a posteriori estimation in dynamical models of primary HIV infection. *Stat. Commun. Infec. Dis.* **4** Art. 2, 36. MR2945221

FINKENSTÄDT, B., WOODCOCK, D. J., KOMOROWSKI, M., HARPER, C. V., DAVIS, J. R. E., WHITE, M. R. H. and RAND, D. A. (2013). Quantifying intrinsic and extrinsic noise in gene transcription using the linear noise approximation: An application to single cell data. *Ann. Appl. Stat.* **7** 1960–1982. MR3161709

FRY, T. J. and MACKALL, C. L. (2002). Interleukin-7: From bench to clinic. *Blood* **99** 3892–3904.

GENZ, A. and KEISTER, B. D. (1996). Fully symmetric interpolatory rules for multiple integrals over infinite regions with Gaussian weight. *J. Comput. Appl. Math.* **71** 299–309. MR1399898

GUEDJ, J., THIÉBAUT, R. and COMMENGES, D. (2007a). Maximum likelihood estimation in dynamical models of HIV. *Biometrics* **63** 1198–1206, 1314. MR2414598

GUEDJ, J., THIÉBAUT, R. and COMMENGES, D. (2007b). Practical identifiability of HIV dynamics models. *Bull. Math. Biol.* **69** 2493–2513. MR2353843

HO, D. D., NEUMANN, A. U., PERELSON, A. S., CHEN, W., LEONARD, J. M., MARKOWITZ, M. et al. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373** 123–126.

HUANG, Y., LIU, D. and WU, H. (2006). Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics* **62** 413–423. MR2227489

KONDRACK, R. M., HARBERTSON, J., TAN, J. T., MCBREEN, M. E., SURH, C. D. and BRADLEY, L. M. (2003). Interleukin 7 regulates the survival and generation of memory CD4 cells. *J. Exp. Med.* **198** 1797–1806.

KONISHI, S. and KITAGAWA, G. (2008). *Information Criteria and Statistical Modeling*. Springer, New York. MR2367855

KUHN, E. and LAVIELLE, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal.* **49** 1020–1038. MR2143055

LEVY, Y., LACABARATZ, C., WEISS, L., VIARD, J.-P., GOUJARD, C., LELIÈVRE, J.-D., BOUÉ, F., MOLINA, J.-M., ROUZIOUX, C., AVETTAND-FÉNOËL, V. et al. (2009). Enhanced T cell recovery in HIV-1-infected adults through IL-7 treatment. *J. Clin. Invest.* **119** 997.

LEVY, Y., SERETI, I., TAMBUSSI, G., ROUTY, J., LELIEVRE, J., DELFRAISSY, J., MOLINA, J., FISCHL, M., GOUJARD, C., RODRIGUEZ, B. et al. (2012). Effects of recombinant human interleukin 7 on T-cell recovery and thymic output in HIV-infected patients receiving antiretroviral therapy: Results of a phase I/IIa randomized, placebo-controlled, multicenter study. *Clin. Infect. Dis.* **55** 291–300.

MACKALL, C. L., FRY, T. J. and GRESS, R. E. (2011). Harnessing the biology of IL-7 for therapeutic application. *Nat. Rev., Immunol.* **11** 330–342.

MACKALL, C. L., FRY, T. J., BARE, C., MORGAN, P., GALBRAITH, A. and GRESS, R. E. (2001). IL-7 increases both thymic-dependent and thymic-independent T-cell regeneration after bone marrow transplantation. *Blood* **97** 1491–1497.

NAMEN, A., SCHMIERER, A., MARCH, C., OVERELL, R., PARK, L., URDAL, D. and MOCHIZUKI, D. (1988). B cell precursor growth-promoting activity. Purification and characterization of a growth factor active on lymphocyte precursors. *J. Exp. Med.* **167** 988–1002.

OKAMOTO, Y., DOUEK, D. C., MCFARLAND, R. D. and KOUP, R. A. (2002). Effects of exogenous interleukin-7 on human thymus function. *Blood* **99** 2851–2858.

PERELSON, A. S., NEUMANN, A. U., MARKOWITZ, M., LEONARD, J. M. and HO, D. D. (1996). HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. *Science* **271** 1582–1586.

PINHEIRO, J. C. and BATES, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

POST, T. M., FREIJER, J. I., PLOEGER, B. A. and DANHOF, M. (2008). Extensions to the visual predictive check to facilitate model performance evaluation. *J. Pharmacokinet. Pharmacodyn.* **35** 185–202.

PRAGUE, M., COMMENGES, D., DRYLEWICZ, J. and THIÉBAUT, R. (2012). Treatment monitoring of HIV-infected patients based on mechanistic models. *Biometrics* **68** 902–911. MR3055195

PRAGUE, M., COMMENGES, D., GUEDJ, J., DRYLEWICZ, J. and THIÉBAUT, R. (2013). NIMROD: A program for inference via a normal approximation of the posterior in models with random effects based on ordinary differential equations. *Comput. Methods Programs Biomed.* **111** 447–458.

RAMSAY, J. O., HOOKER, G., CAMPBELL, D. and CAO, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 741–796. MR2368570

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602

SEDDON, B., TOMLINSON, P. and ZAMOYSKA, R. (2003). Interleukin 7 and T cell receptor signals regulate homeostasis of CD4 memory cells. *Nat. Immunol.* **4** 680–686.

SERETI, I., DUNHAM, R. M., SPRITZLER, J., AGA, E., PROSCHAN, M. A., MEDVIK, K., BATTAGLIA, C. A., LANDAY, A. L., PAHWA, S., FISCHL, M. A. et al. (2009). IL-7 administration drives T cell-cycle entry and expansion in HIV-1 infection. *Blood* **113** 6304–6314.

SPORTÈS, C., HAKIM, F. T., MEMON, S. A., ZHANG, H., CHUA, K. S., BROWN, M. R., FLEISHER, T. A., KRUMLAUF, M. C., BABB, R. R., CHOW, C. K. et al. (2008). Administration of rhIL-7 in humans increases in vivo TCR repertoire diversity by preferential expansion of naive T cell subsets. *J. Exp. Med.* **205** 1701–1714.

THIÉBAUT, R., JACQMIN-GADDA, H., LEPORT, C., KATLAMA, C., COSTAGLIOLA, D., LE MOING, V., MORLAT, P., CHÊNE, G., GROUP, A. S. et al. (2003). Bivariate longitudinal model for the analysis of the evolution of HIV RNA and CD4 cell count in HIV infection taking into account left censoring of HIV RNA measures. *J. Biopharm. Statist.* **13** 271–282.

THIÉBAUT, R., DRYLEWICZ, J., PRAGUE, M., LACABARATZ, C., BEQ, S., JARNE, A., CROUGHS, T., SEKALY, R.-P., LEDERMAN, M. M., SERETI, I. et al. (2014). Quantifying and predicting the effect of exogenous Interleukin-7 on CD4+ T cells in HIV-1 infection. *PLoS Comput. Biol.* **10** e1003630.

THIÉBAUT, R., JARNE, A., ROUTY, J.-P., SERETI, I., FISCHL, M., IVE, P., SPECK, R. et al. (2016). Repeated cycles of recombinant human interleukin 7 in HIV-infected patients with low CD4 T cell reconstitution on antiretroviral therapy: Results of two phase II multicenter studies. *Clin. Infect. Dis.* **62** 1178–1185.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

VIEIRA, M., SOARES, D., BORTHWICK, N. J., MAINI, M. K., JANOSSY, G., SALMON, M. and AKBAR, A. N. (1998). IL-7-dependent extrathymic expansion of CD45RA+ T cells enables preservation of a naive repertoire. *J. Immunol.* **161** 5909–5917.

WANG, L., CAO, J., RAMSAY, J. O., BURGER, D. M., LAPORTE, C. J. L. and ROCKSTROH, J. K. (2014). Estimating mixed-effects differential equation models. *Stat. Comput.* **24** 111–121. MR3147702

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1617–1648
DOI: 10.1214/17-AOAS1049

DYNAMIC MIXTURES OF FACTOR ANALYZERS TO CHARACTERIZE MULTIVARIATE AIR POLLUTANT EXPOSURES

BY ANTONELLO MARUOTTI[*,†,1], JAN BULLA[‡,1], FRANCESCO LAGONA[§,2], MARCO PICONE[¶] AND FRANCESCA MARTELLA[‖]

University of Southampton[*], Libera Università Maria Ss. Assunta[†], University of Bergen[‡], Università di Roma Tre[§], The Institute for Environmental Protection and Research (ISPRA)[¶] and Sapienza Università di Roma[‖]

The assessment of pollution exposure is based on the analysis of a multivariate time series that include the concentrations of several pollutants as well as the measurements of multiple atmospheric variables. It typically requires methods of dimensionality reduction that are capable of identifying potentially dangerous combinations of pollutants and simultaneously segmenting exposure periods according to air quality conditions. When the data are high-dimensional, however, efficient methods of dimensionality reduction are challenging because of the formidable structure of cross-correlations that arise from the dynamic interaction between weather conditions and natural/anthropogenic pollution sources. In order to assess pollution exposure in an urban area while taking the above mentioned difficulties into account, we have developed a class of parsimonious hidden Markov models. In a multivariate time series setting, this approach simultaneously allows for the performance of temporal segmentation and dimensionality reduction. We specifically approximate the distribution of multiple pollutant concentrations by mixtures of factor analysis models, whose parameters evolve according to a latent Markov chain. Covariates are included as predictors of the chain transition probabilities. Parameter constraints on the factorial component of the model are exploited to tune the flexibility of dimensionality reduction. In order to estimate the model parameters efficiently, we have proposed a novel three-step Alternating Expected Conditional Maximization (AECM) algorithm, which is also assessed in a simulation study. In the case study, the proposed methods could (1) describe the exposure to pollution in terms of a few latent regimes, (2) associate these regimes with specific combinations of pollutant concentration levels as well as distinct correlation structures between concentrations, and (3) capture the influence of weather conditions on transitions between regimes.

REFERENCES

ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. MR2549554

ANDERSON, T. W. and RUBIN, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. V* (J. Neyman, ed.) 111–150. Univ. California Press, Berkeley, CA. MR0084943

Key words and phrases. Hidden Markov models, AECM algorithm, dimensionality reduction, three-step algorithm.

BARTOLUCCI, F. and FARCOMENI, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *J. Amer. Statist. Assoc.* **104** 816–831. MR2751454

BARTOLUCCI, F. and FARCOMENI, A. (2010). A note on the mixture transition distribution and hidden Markov models. *J. Time Series Anal.* **31** 132–138. MR2677343

BARTOLUCCI, F., FARCOMENI, A. and PENNONI, F. (2013). *Latent Markov Models for Longitudinal Data*. CRC Press, Boca Raton, FL. MR3184304

BARTOLUCCI, F., MONTANARI, G. E. and PANDOLFI, S. (2015). Three-step estimation of latent Markov models with covariates. *Comput. Statist. Data Anal.* **83** 287–301. MR3281812

BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* **41** 164–171. MR0287613

BORNN, L., SHADDICK, G. and ZIDEK, J. V. (2012). Modeling nonstationary processes through dimension expansion. *J. Amer. Statist. Assoc.* **107** 281–289. MR2949359

BOUVEYRON, C. and BRUNET-SAUMARD, C. (2014). Model-based clustering of high-dimensional data: A review. *Comput. Statist. Data Anal.* **71** 52–78. MR3131954

BULLA, J. and BERZEL, A. (2008). Computational issues in parameter estimation for stationary hidden Markov models. *Comput. Statist.* **23** 1–18. MR2434751

BULLA, J., LAGONA, F., MARUOTTI, A. and PICONE, M. (2012). A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series. *J. Agric. Biol. Environ. Stat.* **17** 544–567. MR3041884

CHATTOPADHYAY, A. K., MONDAL, S. and BISWAS, A. (2015). Independent component analysis and clustering for pollution data. *Environ. Ecol. Stat.* **22** 33–43. MR3316693

CHATZIS, S. P. (2010). Hidden Markov models with nonelliptically contoured state densities. *IEEE Trans. Pattern Anal. Mach. Intell.* **32** 2297–2304.

CHATZIS, S. P., KOSMOPOULOS, D. I. and VARVARIGOU, T. A. (2009). Robust sequential data modelling using an outlier tolerant hidden Markov model. *IEEE Trans. Pattern Anal. Mach. Intell.* **31** 1657–1669.

COOLEY, D., DAVIS, R. A. and NAVEAU, P. (2012). Approximating the conditional density given large observed values via a multivariate extremes framework, with application to environmental data. *Ann. Appl. Stat.* **6** 1406–1429. MR3058669

DANNEMANN, J., HOLZMANN, H. and LESITER, A. (2014). Semiparameteric hidden Markov models: Identifiability and estimation. *Wiley Interdiscip. Rev.: Comput. Stat.* **6** 418–425.

FARCOMENI, A. and GRECO, L. (2015). S-estimation of hidden Markov models. *Comput. Statist.* **30** 57–80. MR3334711

FASSÒ, A., CAMELETTI, M. and NICOLIS, O. (2007). Air quality monitoring using heterogeneous networks. *Environmetrics* **18** 245–264. MR2364347

FIELD, M., STIRLING, D., PAN, Z. and NAGHDY, F. (2016). Learning trajectories for robot programming by demonstration using a coordinated mixture of factor analyzers. *IEEE Trans. Cybern.* **46** 706–717.

GHAHRAMANI, Z. and HINTON, G. E. (1997). The EM algorithm for factor analyzers. Technical report CRG-TR-96-1, Univ. Toronto.

GREVEN, S., DOMINICI, F. and ZEGER, S. (2011). An approach to the estimation of chronic air pollution effects using spatio-temporal information. *J. Amer. Statist. Assoc.* **106** 395–406. MR2866970

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.

KIM, K. H., JAHAN, S. A. and KABIR, E. (2013). A review on human health perspective of air pollution with respect to allergies and asthma. *Environ. Int.* **59** 41–52.

LAGONA, F., MARUOTTI, A. and PADOVANO, F. (2015). Multilevel multivariate modelling of legislative count data, with a hidden Markov chain. *J. Roy. Statist. Soc. Ser. A* **178** 705–723. MR3348355

LAGONA, F., MARUOTTI, A. and PICONE, M. (2011). A non-homogeneous hidden Markov model for the analysis of multi-pollutant exceedances data. In *Hidden Markov Models*: *Theory and Applications* (P. Dymarski, ed.) 207–222. InTech Publisher, Rijeka.

LAGONA, F., PICONE, M. and MARUOTTI, A. (2015). A hidden Markov model for the analysis of cyclindrical time series. *Environmetrics* **26** 534–544. MR3431928

LATZA, U., GERDES, S. and BAUR, X. (2009). Effects of nitrogen dioxide on human health: Systematic review of experimental and epidemiological studies conducted between 2002 and 2006. *Int. J. Hyg. Environ. Health* **212** 271–287.

LEE, D., RUSHWORTH, A. and SAHU, S. K. (2014). A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution. *Biometrics* **70** 419–429. MR3258046

LEE, D. and SAHU, S. (2016). Estimating the health impact of environmental pollution fields. In *Handbook of Spatial Epidemiology* (A. Lawson, S. Banerjee, R. Haining and L. Ugarte, eds.) 271 – 278. Chapman & Hall/CRC, Boca Raton, FL.

LEROUX, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl.* **40** 127–143. MR1145463

MARTINEZ-ZARZOSO, I. and MARUOTTI, A. (2013). The environmental Kuznets curve: Functional form, time-varying heterogeneity and outliers in a panel setting. *Environmetrics* **24** 461–475. MR3137747

MARUOTTI, A. (2011). Mixed hidden Markov models for longitudinal data: An overview. *Int. Stat. Rev.* **79** 427–454.

MARUOTTI, A. and ROCCI, R. (2012). A mixed non-homogeneous hidden Markov model for categorical data, with application to alcohol consumption. *Stat. Med.* **31** 871–886. MR2913866

MARUOTTI, A., PUNZO, A., MASTRANTONIO, G. and LAGONA, F. (2016). A time-dependent extension of the projected normal regression model for longitudinal circular data based on a hidden Markov heterogeneity structure. *Stoch. Environ. Res. Risk Assess.* **30** 1725–1740.

MCLACHLAN, G. J., PEEL, D. and BEAN, R. W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Comput. Statist. Data Anal.* **41** 379–388. MR1973720

MCNICHOLAS, P. D. and MURPHY, T. B. (2008). Parsimonious Gaussian mixture models. *Stat. Comput.* **18** 285–296. MR2413385

MCNICHOLAS, P. D. and MURPHY, T. B. (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics* **26** 2705–2712.

MENG, X.-L. and VAN DYK, D. (1997). The EM algorithm—An old folk-song sung to a fast new tune. *J. Roy. Statist. Soc. Ser. B* **59** 511–567. MR1452025

PACIOREK, C. J., YANOSKY, J. D., PUETT, R. C., LADEN, F. and SUH, H. H. (2009). Practical large-scale spatio-temporal modeling of particulate matter concentrations. *Ann. Appl. Stat.* **3** 370–397. MR2668712

PARK, E. S., GUTTORP, P. and HENRY, R. C. (2001). Multivariate receptor modeling for temporally correlated data by using MCMC. *J. Amer. Statist. Assoc.* **96** 1171–1183. MR1946572

POLLICE, A. and JONA LASINIO, G. (2009). Two approaches to imputation and adjustment of air quality data from a composite monitoring network. *J. Data Sci.* **7** 43–59.

PUNZO, A. and MARUOTTI, A. (2016). Clustering multivariate longitudinal observations: The contaminated Gaussian hidden Markov model. *J. Comput. Graph. Statist.* **25** 1097–1116. MR3572030

PUNZO, A. and MCNICHOLAS, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biom. J.* **58** 1506–1537.

ROSTI, A. V. I. and GALES, M. J. F. (2002). Factor analysed hidden Markov models. In *Proceedings of the IEEE International Conference on Acoustics*, *Speech and Signal Processing* 949–952.

SAHU, S. K., BAFFOUR, B., HARPER, P. R., MINTY, J. H. and SARRAN, C. (2014). A hierarchical Bayesian model for improving short-term forecasting of hospital demand by including meteorological information. *J. Roy. Statist. Soc. Ser. A* **177** 39–61. MR3158666

SCOTT, S. L., JAMES, G. M. and SUGAR, C. A. (2005). Hidden Markov models for longitudinal comparisons. *J. Amer. Statist. Assoc.* **100** 359–369. MR2170459

SHADDICK, G. and WAKEFIELD, J. (2002). Modelling daily multivariate pollutant data at multiple sites. *J. Roy. Statist. Soc. Ser. C* **51** 351–372. MR1920802

SHADDICK, G., LEE, D., ZIDEK, J. V. and SALWAY, R. (2008). Estimating exposure response functions using ambient pollution concentrations. *Ann. Appl. Stat.* **2** 1249–1270. MR2655658

VISSER, I., RAIJMAKERS, M. and MOLENAAR, P. (2000). Confidence intervals for hidden Markov model parameters. *Br. J. Math. Stat. Psychol.* **53** 317–327.

VITERBI, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform. Theory* **13** 260–269.

WELCH, L. (2003). Hidden Markov models and the Baum–Welch algorithm. *IEEE Inf. Theory Soc. Newsl.* **53** 1–13.

YAO, K., PALIWAL, K. K. and LEE, T. W. (2005). Generative factor analyzed HMM for automatic speech recognition. *Speech Commun.* **45** 435–454.

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

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1649–1670
DOI: 10.1214/17-AOAS1050

DYNAMIC PREDICTION OF DISEASE PROGRESSION FOR LEUKEMIA PATIENTS BY FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS OF LONGITUDINAL EXPRESSION LEVELS OF AN ONCOGENE

BY FANGRONG YAN[*,†,1], XIAO LIN[*] AND XUELIN HUANG[†,2]

China Pharmaceutical University[*] and The University of Texas MD Anderson Cancer Center[†]

Patients' biomarker data are repeatedly measured over time during their follow-up visits. Statistical models are needed to predict disease progression on the basis of these longitudinal biomarker data. Such predictions must be conducted on a real-time basis so that at any time a new biomarker measurement is obtained, the prediction can be updated immediately to reflect the patient's latest prognosis and further treatment can be initiated as necessary. This is called dynamic prediction. The challenge is that longitudinal biomarker values fluctuate over time, and their changing patterns vary greatly across patients. In this article, we apply functional principal components analysis (FPCA) to longitudinal biomarker data to extract their features, and use these features as covariates in a Cox proportional hazards model to conduct dynamic predictions. Our flexible approach comprehensively characterizes the trajectory patterns of the longitudinal biomarker data. Simulation studies demonstrate its robust performance for dynamic prediction under various scenarios. The proposed method is applied to dynamically predict the risk of disease progression for patients with chronic myeloid leukemia following their treatments with tyrosine kinase inhibitors. The FPCA method is applied to their longitudinal measurements of *BCR-ABL* gene expression levels during follow-up visits to obtain the changing patterns over time as predictors.

REFERENCES

ANTOLINI, L., BORACCHI, P. and BIGANZOLI, E. (2005). A time-dependent discrimination index for survival data. *Stat. Med.* **24** 3927–3944. MR2221976

BERKEY, C. and KENT, R. J. (2009). Longitudinal principal components and non-linear regression models of early childhood growth. *Ann. Hum. Biol.* **10** 523–536.

BESSE, P. and RAMSAY, J. O. (1986). Principal components analysis of sampled functions. *Psychometrika* **51** 285–311. MR0848110

BRESLOW, N. E. (1972). Discussion of "Regression models and life-tables" by D. R. Cox. *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

BROWN, E. R., IBRAHIM, J. G. and DEGRUTTOLA, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* **61** 64–73. MR2129202

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

Key words and phrases. Dynamic prediction, functional principal component analysis, longitudinal biomarker, joint modeling, survival analysis.

DAI, X., HADJIPANTELIS, P. Z., JI, H., MUELLER, H. G. and WANG, J. L. (2016). Functional data analysis and empirical dynamics. Available at https://cran.r-project.org/web/packages/fdapace/fdapace.pdf.

GRAMBSCH, P. M. and THERNEAU, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81** 515–526. MR1311094

GRANT, S., CHEN, Y. Q. and MAY, S. (2014). Performance of goodness-of-fit tests for the Cox proportional hazards model with time-varying covariates. *Lifetime Data Anal.* **20** 355–368. MR3217542

HALL, P., MÜLLER, H.-G. and WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. MR2278365

HARRELL, F. E., LEE, K. L. and MARK, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15** 361–387.

HEAGERTY, P. J. and ZHENG, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61** 92–105. MR2135849

HUANG, X. and LIU, L. (2007). A joint frailty model for survival and gap times between recurrent events. *Biometrics* **63** 389–397. MR2370797

HUANG, X., YAN, F., NING, J., FENG, Z., CHOI, S. and CORTES, J. (2016). A two-stage approach for dynamic prediction of time-to-event distributions. *Stat. Med.* **35** 2167–2182. MR3513506

IBRAHIM, J. G., CHEN, M.-H. and SINHA, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statist. Sinica* **14** 863–883. MR2087976

JAMES, G. M., HASTIE, T. J. and SUGAR, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87** 587–602. MR1789811

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481. MR0093867

LENG, X. and MÜLLER, H.-G. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics* **22** 68–76.

LIN, D. Y., WEI, L. J. and YING, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80** 557–572. MR1248021

LIN, J., ZHANG, D. and DAVIDIAN, M. (2006). Smoothing spline-based score tests for proportional hazards models. *Biometrics* **62** 803–812. MR2247209

LIU, L. and HUANG, X. (2009). Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 65–81. MR2662234

LIU, X. and YANG, M. C. K. (2009). Identifying temporally differentially expressed genes through functional principal components analysis. *Biostatistics* **10** 667–679.

PAULER, D. and FINKELSTEIN, D. (2002). Predicting time to prostate cancer recurrence based on joint models for non-linear longitudinal biomarkers and event time outcomes. *Stat. Med.* **21(24)** 3897–3911.

QUINTAS-CARDAMA, A., CHOI, S., KANTARJIAN, H., JABBOUR, E., HUANG, X. and CORTES, J. (2014). Predicting outcomes in patients with chronic myeloid leukemia at any time during tyrosine kinase inhibitor therapy. *Clin. Lymphoma Myeloma Leuk.* **14** 327–334.

RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53** 233–243. MR1094283

RIZOPOULOS, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *J. Stat. Softw.* **35 (9)** 1–33.

RIZOPOULOS, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67** 819–829. MR2829256

RIZOPOULOS, D. and GHOSH, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat. Med.* **30** 1366–1380. MR2828959

RIZOPOULOS, D., HATFIELD, L. A., CARLIN, B. P. and TAKKENBERG, J. J. M. (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *J. Amer. Statist. Assoc.* **109** 1385–1397. MR3293598

SILVERMAN, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Statist.* **24** 1–24. MR1389877

SLATE, E. and TURNBULL, B. (2000). Statistical models for longitudinal biomarkers of disease onset. *Stat. Med.* **19(4)** 617–637.

SONG, X., DAVIDIAN, M. and TSIATIS, A. A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* **58** 742–753. MR1945011

STANISWALIS, J. G. and LEE, J. J. (1998). Nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc.* **93** 1403–1418. MR1666636

TSIATIS, A. A. and DAVIDIAN, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88** 447–458. MR1844844

UNO, H., CAI, T., TIAN, L. and WEI, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *J. Amer. Statist. Assoc.* **102** 527–537. MR2370850

WULFSOHN, M. S. and TSIATIS, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53** 330–339. MR1450186

XU, J. and ZEGER, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *J. Roy. Statist. Soc. Ser. C* **50** 375–387. MR1856332

YAO, F. and LEE, T. C. M. (2006). Penalized spline models for functional principal component analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 3–25. MR2212572

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561

YAO, F., MÜLLER, H.-G., CLIFFORD, A. J., DUEKER, S. R., FOLLETT, J., LIN, Y., BUCHHOLZ, B. A. and VOGEL, J. S. (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics* **59** 676–685. MR2004273

ZHENG, Y., CAI, T. and FENG, Z. (2006). Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics* **62** 279–287, 321. MR2226583

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1671–1710
DOI: 10.1214/17-AOAS1053

SHAPE-CONSTRAINED UNCERTAINTY QUANTIFICATION IN UNFOLDING STEEPLY FALLING ELEMENTARY PARTICLE SPECTRA

BY MIKAEL KUUSELA[1,2] AND PHILIP B. STARK

École Polytechnique Fédérale de Lausanne and University of California, Berkeley

The high energy physics unfolding problem is an important statistical inverse problem in data analysis at the Large Hadron Collider (LHC) at CERN. The goal of unfolding is to make nonparametric inferences about a particle spectrum from measurements smeared by the finite resolution of the particle detectors. Previous unfolding methods use ad hoc discretization and regularization, resulting in confidence intervals that can have significantly lower coverage than their nominal level. Instead of regularizing using a roughness penalty or stopping iterative methods early, we impose physically motivated shape constraints: positivity, monotonicity, and convexity. We quantify the uncertainty by constructing a nonparametric confidence set for the true spectrum, consisting of all those spectra that satisfy the shape constraints and that predict the observations within an appropriately calibrated level of fit. Projecting that set produces simultaneous confidence intervals for all functionals of the spectrum, including averages within bins. The confidence intervals have guaranteed conservative frequentist finite-sample coverage in the important and challenging class of unfolding problems for steeply falling particle spectra. We demonstrate the method using simulations that mimic unfolding the inclusive jet transverse momentum spectrum at the LHC. The shape-constrained intervals provide usefully tight conservative inferences, while the conventional methods suffer from severe undercoverage.

REFERENCES

ADYE, T. (2011). Unfolding algorithms and tests using RooUnfold. In *Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding* (H. B. Prosper and L. Lyons, eds.). *CERN*-2011-006 313–318.

ANTONIADIS, A. and BIGOT, J. (2006). Poisson inverse problems. *Ann. Statist.* **34** 2132–2158. MR2291495

ATLAS COLLABORATION (2012). Measurement of the transverse momentum distribution of W bosons in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Phys. Rev. D* **85** 012005.

BACKUS, G. (1970). Inference from inadequate and inaccurate data, I. *Proc. Natl. Acad. Sci. USA* **65** 1–7. MR0254943

BANERJEE, M. and WELLNER, J. A. (2001). Likelihood ratio tests for monotone functions. *Ann. Statist.* **29** 1699–1731. MR1891743

BARNEY, D. (2004). CMS-doc-4172. Available at https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=4172. Retrieved 21.1.2014.

Key words and phrases. Poisson inverse problem, finite-sample coverage, high energy physics, Large Hadron Collider, Fenchel duality, semi-infinite programming.

BLOBEL, V. (2013). Unfolding. In *Data Analysis in High Energy Physics*: *A Practical Guide to Statistical Methods* (O. Behnke, K. Kröninger, G. Schott and T. Schörner-Sadenius, eds.) 187–225. Wiley, Weinheim.

BURRUS, W. R. (1965). Utilization of a priori information by means of mathematical programming in the statistical interpretation of measured distributions. ORNL-3743, Oak Ridge National Laboratory.

BURRUS, W. R. and VERBINSKI, V. V. (1969). Fast-neutron spectroscopy with thick organic scintillators. *Nucl. Instrum. Methods* **67** 181–196.

CAI, T. T., LOW, M. G. and XIA, Y. (2013). Adaptive confidence intervals for regression functions under shape constraints. *Ann. Statist.* **41** 722–750. MR3099119

CARROLL, R. J., DELAIGLE, A. and HALL, P. (2011). Testing and estimating shape-constrained nonparametric density and regression in the presence of measurement error. *J. Amer. Statist. Assoc.* **106** 191–202. MR2816713

CHOUDALAKIS, G. (2012). Fully Bayesian unfolding. Preprint. Available at arXiv:1201.4612v4 [physics.data-an].

CMS COLLABORATION (2008). The CMS experiment at the CERN LHC. *J. Instrum.* **3** S08004.

CMS COLLABORATION (2010). Measurement of the inclusive jet cross section in pp collisions at 7 TeV. CMS-PAS-QCD-10-011. Available at http://cds.cern.ch/record/1280682.

CMS COLLABORATION (2011). Measurement of the inclusive jet cross section in pp collisions at $\sqrt{s} = 7$ TeV. *Phys. Rev. Lett.* **107** 132001.

CMS COLLABORATION (2013a). Measurements of differential jet cross sections in proton-proton collisions at $\sqrt{s} = 7$ TeV with the CMS detector. *Phys. Rev. D* **87** 112002.

CMS COLLABORATION (2013b). Measurement of differential top-quark-pair production cross sections in pp collisions at $\sqrt{s} = 7$ TeV. *Eur. Phys. J. C* **73** 2339.

CMS COLLABORATION (2016). Measurement of differential cross sections for Higgs boson production in the diphoton decay channel in pp collisions at $\sqrt{s} = 8$ TeV. *Eur. Phys. J. C* **76** 13.

COWAN, G. (1998). *Statistical Data Analysis*. Oxford Univ. Press, London.

D'AGOSTINI, G. (1995). A multidimensional unfolding method based on Bayes' theorem. *Nucl. Instrum. Methods Phys. Res., Sect. A* **362** 487–498.

DAVIES, P. L., KOVAC, A. and MEISE, M. (2009). Nonparametric regression, confidence regions and regularization. *Ann. Statist.* **37** 2597–2625. MR2541440

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DÜMBGEN, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *Ann. Statist.* **26** 288–314. MR1611768

DÜMBGEN, L. (2003). Optimal confidence bands for shape-restricted curves. *Bernoulli* **9** 423–449. MR1997491

FORTE, S. and WATT, G. (2013). Progress in the determination of the partonic structure of the proton. *Annu. Rev. Nucl. Part. Sci.* **63** 291–328.

GARWOOD, F. (1936). Fiducial limits for the Poisson distribution. *Biometrika* **28** 437–442.

GENOVESE, C. and WASSERMAN, L. (2008). Adaptive confidence bands. *Ann. Statist.* **36** 875–905. MR2396818

GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*: *A Roughness Penalty Approach*. *Monographs on Statistics and Applied Probability* **58**. Chapman & Hall, London. MR1270012

GRENANDER, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153. MR0093415

GROENEBOOM, P. and JONGBLOED, G. (2014). *Nonparametric Estimation Under Shape Constraints*: *Estimators, Algorithms and Asymptotics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge Univ. Press, New York. MR3445293

GROENEBOOM, P. and JONGBLOED, G. (2015). Nonparametric confidence intervals for monotone functions. *Ann. Statist.* **43** 2019–2054. MR3375875

GROENEBOOM, P., JONGBLOED, G. and WELLNER, J. A. (2001). Estimation of a convex function: Characterizations and asymptotic theory. *Ann. Statist.* **29** 1653–1698. MR1891742

HALL, P. and HOROWITZ, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *Ann. Statist.* **41** 1892–1921. MR3127852

HANSEN, P. C. (1998). *Rank-Deficient and Discrete Ill-Posed Problems*: *Numerical Aspects of Linear Inversion*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. MR1486577

HANSEN, P. C. (2010). *Discrete Inverse Problems*: *Insight and Algorithms*. *Fundamentals of Algorithms* **7**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. MR2584074

HENGARTNER, N. W. and STARK, P. B. (1992). Conservative finite-sample confidence envelopes for monotone and unimodal densities. Technical Report No. 341, Dept. Statistics, Univ. California, Berkeley.

HENGARTNER, N. W. and STARK, P. B. (1995). Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* **23** 525–550. MR1332580

HÖCKER, A. and KARTVELISHVILI, V. (1996). SVD approach to data unfolding. *Nucl. Instrum. Methods Phys. Res., Sect. A* **372** 469–481.

KONDOR, A. (1983). Method of convergent weights—An iterative procedure for solving Fredholm's integral equations of the first kind. *Nucl. Instrum. Methods* **216** 177–181.

KUUSELA, M. and PANARETOS, V. M. (2015). Statistical unfolding of elementary particle spectra: Empirical Bayes estimation and bias-corrected uncertainty quantification. *Ann. Appl. Stat.* **9** 1671–1705. MR3418740

LANGE, K. and CARSON, R. (1984). EM reconstruction algorithms for emission and transmission tomography. *J. Comput. Assist. Tomogr.* **8** 306–316.

LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. MR1604412

LUCY, L. B. (1974). An iterative technique for the rectification of observed distributions. *Astron. J.* **79** 745–754.

LUENBERGER, D. G. (1969). *Optimization by Vector Space Methods*. Wiley, New York. MR0238472

Mathworks (2014). Optimization Toolbox User's Guide. Release 2014a.

MEISTER, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. *Lecture Notes in Statistics* **193**. Springer, Berlin. MR2768576

MÜLTHEI, H. N. and SCHORR, B. (1987a). On an iterative method for a class of integral equations of the first kind. *Math. Methods Appl. Sci.* **9** 137–168. MR0897263

MÜLTHEI, H. N. and SCHORR, B. (1987b). On an iterative method for the unfolding of spectra. *Nucl. Instrum. Methods Phys. Res., Sect. A* **257** 371–377.

MÜLTHEI, H. N. and SCHORR, B. (1989). On properties of the iterative maximum likelihood reconstruction method. *Math. Methods Appl. Sci.* **11** 331–342. MR0991270

NNPDF COLLABORATION (2015). Parton distributions for the LHC run II. *J. High Energy Phys.* **1504** 040.

O'LEARY, D. P. and RUST, B. W. (1986). Confidence intervals for inequality-constrained least squares problems, with applications to ill-posed problems. *SIAM J. Sci. Statist. Comput.* **7** 473–489. MR0833916

O'SULLIVAN, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.* **1** 502–527. MR0874480

PFLUG, G. CH. and WETS, R. J.-B. (2013). Shape-restricted nonparametric regression with overall noisy measurements. *J. Nonparametr. Stat.* **25** 323–338. MR3056088

PHILLIPS, D. L. (1962). A technique for the numerical solution of certain integral equations of the first kind. *J. ACM* **9** 84–97. MR0134481

PIERCE, J. E. and RUST, B. W. (1985). Constrained least squares interval estimation. *SIAM J. Sci. Statist. Comput.* **6** 670–673. MR0791192

PROSPER, H. B. and LYONS, L., eds. (2011). *Proceedings of the PHYSTAT* 2011 *Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*. *CERN*-2011-006.

REISS, R.-D. (1993). *A Course on Point Processes*. Springer, New York. MR1199815

RICHARDSON, W. H. (1972). Bayesian-based iterative method of image restoration. *J. Opt. Soc. Amer.* **62** 55–59.

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, Chichester. MR0961262

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

RUST, B. W. and BURRUS, W. R. (1972). *Mathematical Programming and the Numerical Solution of Linear Equations*. American Elsevier, Publishing Co., Inc., New York. MR0353644

RUST, B. W. and O'LEARY, D. P. (1994). Confidence intervals for discrete approximations to ill-posed problems. *J. Comput. Graph. Statist.* **3** 67–96. MR1273034

SCHMITT, S. (2012). TUnfold, an algorithm for correcting migration effects in high energy physics. *J. Instrum.* **7** T10003.

SHEPP, L. A. and VARDI, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Med. Imag.* **1** 113–122.

STARK, P. B. (1992). Inference in infinite-dimensional inverse problems: Discretization and duality. *J. Geophys. Res.* **97** 14055–14082.

TIKHONOV, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Sov. Math., Dokl.* **4** 1035–1038.

VARDI, Y., SHEPP, L. A. and KAUFMAN, L. (1985). A statistical model for positron emission tomography. *J. Amer. Statist. Assoc.* **80** 8–37. MR0786595

VOGEL, C. R. (2002). *Computational Methods for Inverse Problems*. *Frontiers in Applied Mathematics* **23**. SIAM, Philadelphia, PA. MR1928831

VOLOBOUEV, I. (2015). On the expectation-maximization unfolding with smoothing. Preprint. Available at arXiv:1408.6500v2 [physics.data-an].

VOUTILAINEN, M. (2012). Personal communication.

WAHBA, G. (1982). Constrained regularization for ill-posed linear operator equations, with applications in meteorology and medicine. In *Statistical Decision Theory and Related Topics*, *III*, *Vol.* 2 (*West Lafayette*, *Ind.*, 1981) 383–418. Academic Press, New York. MR0705326

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1711–1737
DOI: 10.1214/17-AOAS1054

TOWARD BAYESIAN INFERENCE OF THE SPATIAL DISTRIBUTION OF PROTEINS FROM THREE-CUBE FÖRSTER RESONANCE ENERGY TRANSFER DATA[1]

BY JAN-OTTO HOOGHOUDT*, MARGARIDA BARROSO† AND RASMUS WAAGEPETERSEN*

Aalborg University and Albany Medical College†*

Förster resonance energy transfer (FRET) is a quantum-physical phenomenon where energy may be transferred from one molecule to a neighbor molecule if the molecules are close enough. Using fluorophore molecule marking of proteins in a cell, it is possible to measure in microscopic images to what extent FRET takes place between the fluorophores. This provides indirect information of the spatial distribution of the proteins. Questions of particular interest are whether (and if so to which extent) proteins of possibly different types interact or whether they appear independently of each other. In this paper we propose a new likelihood-based approach to statistical inference for FRET microscopic data. The likelihood function is obtained from a detailed modeling of the FRET data-generating mechanism conditional on a protein configuration. We next follow a Bayesian approach and introduce a spatial point process prior model for the protein configurations depending on hyperparameters quantifying the intensity of the point process. Posterior distributions are evaluated using Markov chain Monte Carlo. We propose to infer microscope-related parameters in an initial step from reference data without interaction between the proteins. The new methodology is applied to simulated and real datasets.

REFERENCES

ALBER, F., DOKUDOVSKAYA, S., VEENHOFF, L. M., ZHANG, W., KIPPER, J., DEVOS, D., SUPRAPTO, A., KARNI-SCHMIDT, O., WILLIAMS, R., CHAIT, B. T., ROUT, M. P. and SALI, A. (2017). Determining the architectures of macromolecular assemblies. *Nature* **450** 683–694.

BERNEY, C. and DANUSER, G. (2003). FRET or no FRET: A quantitative comparison. *Biophys. J.* **84** 3992–4010.

BONOMI, M., PELLARIN, R., KIM, S. J., RUSSEL, D., SUNDIN, B. A., RIFFLE, M., JASCHOB, D., RAMSDEN, R., DAVIS, T. N., MULLER, E. G. D. and SALI, A. (2014). Determining protein complex structures based on a Bayesian model of in vivo Förster resonance energy transfer data. *Mol. Cell. Proteomics* **13** 2812–2823. DOI:10.1074/mcp.M114.040824.

BUNT, G. and WOUTERS, F. S. (2004). Visualization of molecular activities inside living cells with fluorescent labels. *Int. Rev. Cytol.* **237** 205–277.

CHEN, H., PUHL III, H. L. and IKEDA, S. R. (2007). Estimating protein–protein interaction affinity in living cells using quantitative Förster resonance energy transfer measurements. *J. Biomed. Opt.* **12** Art. ID 054011.

Key words and phrases. Bayesian inference, Markov chain Monte Carlo, Förster resonance energy transfer, spatial point process, spatial distribution, proteins, fluorophores.

CHEN, L.-C., LLOYD III, W. R., CHANG, C.-W., SUD, D. and MYCEK, M.-A. (2013). Fluorescence lifetime imaging microscopy for quantitative biological imaging. In *Digital Microscopy* (G. Sluder and D. E. Wolf, eds.). *Methods in Cell Biology* **114** 457–488. Academic Press, San Diego, CA.

CLEGG, R. M. (1995). Fluorescence resonance energy transfer. *Curr. Opin. Biotechnol.* **6** 103–110.

CLEGG, R. M. (2006). The history of FRET. In *Reviews in Fluorescence* 2006 (C. D. Geddes and J. R. Lakowicz, eds.) 1–45. Springer, New York.

CORRY, B., JAYATILAKA, D. and RIGBY, P. (2005). A flexible approach to the calculation of resonance energy transfer efficiency between multiple donors and acceptors in complex geometries. *Biophys. J.* **89** 3822–3836.

ELANGOVAN, M., WALLRABE, H., CHEN, Y., DAY, R. N., BARROSO, M. and PERIASAMY, A. (2003). Characterization of one- and two-photon excitation fluorescence resonance energy transfer microscopy. *Methods* **29** 58–73.

ERICKSON, H. P. (2009). Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. *Biol. Proced. Online* **11** 32–51.

FÖRSTER, TH. (1948). Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys.* **437** 55–75. DOI:10.1002/andp.19484370105.

FREDERIX, P. L., DE BEER, E. L., HAMELINK, W. and GERRITSEN, H. C. (2002). Dynamic Monte Carlo simulations to model FRET and photobleaching in systems with multiple donor–acceptor interactions. *J. Phys. Chem. B* **106** 6793–6801.

GAMERMAN, D. and LOPES, H. F. (2006). *Markov Chain Monte Carlo*: *Stochastic Simulation for Bayesian Inference*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2260716

GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J., eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London. MR1397966

GOSWAMI, D., GOWRISHANKAR, K., BILGRAMI, S., GHOSH, S., RAGHUPATHY, R., CHADDA, R., VISHWAKARMA, R., RAO, M. and MAYOR, S. (2008). Nanoclusters of GPI-anchored proteins are formed by cortical Actin-driven activity. *Cell* **135** 1085–1097.

GRYCZYNSKI, Z., GRYCZYNSKI, I. and LAKOWICZ, J. R. (2005). Basics of fluorescence and FRET. In *Molecular Imaging*: *FRET Microscopy and Spectroscopy* 21–56. Elsevier, Amsterdam.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

HEITLER, W. (1954). *The Quantum Theory of Radiation*. Dover, New York.

HOOGHOUDT, J.-O., BARROSO, M. and WAAGEPETERSEN, R. (2017). Supplement A: Preliminary statistical analysis of the three-cube FRET dataset. DOI:10.1214/17-AOAS1054SUPPA.

HOOGHOUDT, J.-O. and WAAGEPETERSEN, R. (2017a). Supplement B: The MCMC sampler. DOI:10.1214/17-AOAS1054SUPPB.

HOOGHOUDT, J.-O. and WAAGEPETERSEN, R. (2017b). Supplement C: Inference of the microscope parameters. DOI:10.1214/17-AOAS1054SUPPC.

KENWORTHY, A. K. (2001). Imaging protein–protein interactions using fluorescence resonance energy transfer microscopy. *Methods* **24** 289–296.

KENWORTHY, A. K. and EDIDIN, M. (1998). Distribution of a Glycosylphosphatidylinositol-anchored protein at the apical surface of MDCK cells examined at a resolution of <100 Å using imaging fluorescence resonance energy transfer. *J. Cell Biol.* **142** 69–84.

KRISSINEL, E. and HENRICK, K. (2007). Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372** 774–797.

LAKOWICZ, J. R. (2009). *Principles of Fluorescence Spectroscopy*. Springer, New York.

LOURA, L. M. S., FERNANDES, F. and PRIETO, M. (2010). Membrane microheterogeneity: Förster resonance energy transfer characterization of lateral membrane domains. *Eur. Biophys. J.* **39** 589–607.

LOURA, L. M. S. and PRIETO, M. (2011). FRET in membrane biophysics: An overview. *Front. Physiol.* **2** Art. ID 82.

MIYAWAKI, A., SAWANO, A. and KOGURE, T. (2003). Lighting up cells: Labelling proteins with fluorophores. *Nat. Cell Biol.* **5** S1–S7.

MØLLER, J. and WAAGEPETERSEN, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes. Monographs on Statistics and Applied Probability* **100**. Chapman & Hall/CRC, Boca Raton, FL. MR2004226

MØLLER, J., PETTITT, A. N., REEVES, R. and BERTHELSEN, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93** 451–458. MR2278096

MURRAY, I., GHAHRAMANI, Z. and MACKAY, D. J. C. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)* 359–366. AUAI Press, Arlington, VA.

PAWLEY, J. (2006). *Handbook of Biological Confocal Microscopy*. Springer, New York.

PERIASAMY, A. and DAY, R., eds. (2011). *Molecular Imaging: FRET Microscopy and Spectroscopy*. Elsevier, Amsterdam.

PERIASAMY, A., WALLRABE, H., CHEN, Y. and BARROSO, M. (2008). Quantitation of protein–protein interactions: Confocal FRET microscopy. In *Biophysical Tools for Biologists, Volume Two: In Vivo Techniques* (J. J. Correia and H. W. I. Detrich, eds.). *Methods in Cell Biology* **89** 569–598. Academic Press, San Diego, CA.

POLO, S. E. and JACKSON, S. P. (2011). Dynamics of DNA damage response proteins at DNA breaks: A focus on protein modifications. *Genes Dev.* **25** 409–433.

PUGLISI, J. D. (2005). *Structure, Dynamics and Function of Biological Macromolecules and Assemblies. NATO Science Series, I: Life and Behavioural Sciences* **364**. IOS Press, Amsterdam.

RAICU, V., STONEMAN, M. R., FUNG, R., MELNICHUK, M., JANSMA, D. B., PISTERZI, L. F., RATH, S., FOX, M., WELLS, J. W. and SALDIN, D. K. (2009). Determination of supramolecular structure and spatial distribution of protein complexes in living cells. *Nat. Photonics* **3** 107–113.

ROHATGI-MUKHERJEE, K. K. (1978). *Fundamentals of Photochemistry*. Wiley, New Delhi.

SHIMA, S. and SAKAI, H. (1977). Polylysine produced by *Streptomyces*. *Agric. Biol. Chem.* **41** 1807–1809.

SUN, Y., WALLRABE, H., SEO, S.-A. and PERIASAMY, A. (2011). FRET microscopy in 2010: The legacy of Theodor Förster on the 100th anniversary of his birth. *Chemphyschem.* **12** 462–474. DOI:10.1002/cphc.201000664.

VAN PUTTEN, E. G., AKBULUT, D., BERTOLOTTI, J., VOS, W. L., LAGENDIJK, A. and MOSK, A. P. (2011). Scattering lens resolves sub-100 nm structures with visible light. *Phys. Rev. Lett.* **106** Art. ID 193905.

WALLRABE, H. and PERIASAMY, A. (2005). Imaging protein molecules using FRET and FLIM microscopy. *Curr. Opin. Biotechnol.* **16** 19–27.

WALLRABE, H., ELANGOVAN, M., BURCHARD, A., PERIASAMY, A. and BARROSO, M. (2003). Confocal FRET microscopy to measure clustering of ligand–receptor complexes in endocytic membranes. *Biophys. J.* **85** 559–571.

WALLRABE, H., CHEN, Y., PERIASAMY, A. and BARROSO, M. (2006). Issues in confocal microscopy for quantitative FRET analysis. *Microsc. Res. Tech.* **69** 196–206.

WALLRABE, H., BONAMY, G., PERIASAMY, A. and BARROSO, M. (2007). Receptor complexes cotransported via polarized endocytic pathways form clusters with distinct organizations. *Mol. Biol. Cell* **18** 2226–2243.

WELCH, S. (1992). *Transferrin: The Iron Carrier*. CRC Press, Boca Raton, FL.

WOLBER, P. K. and HUDSON, B. S. (1979). An analytic solution to the Förster energy transfer problem in two dimensions. *Biophys. J.* **28** 197–210.

WU, P. G. and BRAND, L. (1994). Resonance energy transfer: Methods and applications. *Anal. Biochem.* **218** 1–13.

ZAL, T. and GASCOIGNE, N. R. J. (2004). Photobleaching-corrected FRET efficiency imaging of live cells. *Biophys. J.* **86** 3923–3939.

ZAL, T., ZAL, M. A. and GASCOIGNE, N. R. J. (2002). Inhibition of T cell receptor–coreceptor interactions by antagonist ligands visualized by live FRET imaging of the T-hybridoma immunological synapse. *Immunity* **16** 521–534.

ZIMMERMANN, T., RIETDORF, J. and PEPPERKOK, R. (2003). Spectral imaging and its applications in live cell microscopy. *FEBS Lett*. **546** 87–92.

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1738–1762
DOI: 10.1214/17-AOAS1056

BIOMARKER CHANGE-POINT ESTIMATION WITH RIGHT CENSORING IN LONGITUDINAL STUDIES[1]

BY XIAOYING TANG[*], MICHAEL I. MILLER[†] AND LAURENT YOUNES[†]

Sun Yat-Sen University[*] and Johns Hopkins University[†]

We consider in this paper a statistical two-phase regression model in which the change point of a disease biomarker is measured relative to another point in time, such as the manifestation of the disease, which is subject to right-censoring (i.e., possibly unobserved over the entire course of the study). We develop point estimation methods for this model, based on maximum likelihood, and bootstrap validation methods. The effectiveness of our approach is illustrated by numerical simulations, and by the estimation of a change point for amygdalar atrophy in the context of Alzheimer's disease, wherein it is related to the cognitive manifestation of the disease.

REFERENCES

ALZHEIMER'S ASSOCIATION (2015). 2015 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **11** 332–384.

ATIYA, M., HYMAN, B. T., ALBERT, M. and KILLIANY, R. (2003). Structural magnetic resonance imaging in established and prodromal Alzheimer disease: A review. *Alzheimer Dis. Assoc. Disord.* **17** 177–195.

BAUER, M., BRUVERIS, M. and MICHOR, P. W. (2014). Overview of the geometries of shape spaces and diffeomorphism groups. *J. Math. Imaging Vision* **50** 60–97. MR3233135

CAVEDO, E., BOCCARDI, M., GANZOLA, R., CANU, E., BELTRAMELLO, A., CALTAGIRONE, C., THOMPSON, P. M. and FRISONI, G. B. (2011). Local amygdala structural differences with 3T MRI in patients with Alzheimer disease. *Neurology* **76** 727–733.

CHEN, J. and GUPTA, A. K. (2000). *Parametric Statistical Change Point Analysis*. Birkhäuser, Boston, MA. MR1761850

CSERNANSKY, J. G., WANG, L., SWANK, J., MILLER, J. P., GADO, M., MCKEEL, D., MILLER, M. I. and MORRIS, J. C. (2005). Preclinical detection of Alzheimer's disease: Hippocampal shape and volume predict dementia onset in the elderly. *NeuroImage* **25** 783–792.

DEN HEIJER, T., GEERLINGS, M. I., HOEBEEK, F. E., HOFMAN, A., KOUDSTAAL, P. J. and BRETELER, M. M. B. (2006). Use of hippocampal and amygdalar volumes on magnetic resonance imaging to predict dementia in cognitively intact elderly people. *Arch. Gen. Psychiatry* **63** 57–62.

DUPUY, J.-F. (2006). Estimation in a change-point hazard regression model. *Statist. Probab. Lett.* **76** 182–190. MR2233390

FARLEY, J. U. and HINICH, M. J. (1970). A test for a shifting slope coefficient in a linear model. *J. Amer. Statist. Assoc.* **65** 1320–1329.

FEDER, P. I. (1975). On asymptotic distribution theory in segmented regression problems—Identified case. *Ann. Statist.* **3** 49–83. MR0378267

FISCHL, B. (2012). FreeSurfer. *NeuroImage* **62** 774–781.

Key words and phrases. Change-point estimation, right censoring, medical imaging.

GOMBAY, E. and HORVÁTH, L. (1994a). Limit theorems for change in linear regression. *J. Multivariate Anal.* **48** 43–69. MR1256834

GOMBAY, E. and HORVÁTH, L. (1994b). An application of the maximum likelihood test to the change-point problem. *Stochastic Process. Appl.* **50** 161–171. MR1262337

HAMANN, S. (2001). Cognitive and neural mechanisms of emotional memory. *Trends Cogn. Sci.* **5** 394–400.

HEBERT, L. E., WEUVE, J., SCHERR, P. A. and EVANS, D. A. (2013). Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology* **80** 1778–1783.

HINKLEY, D. V. (1969). Inference about the intersection in two-phase regression. *Biometrika* **56** 495–504.

HINKLEY, D. V. (1971). Inference in two-phase regression. *J. Amer. Statist. Assoc.* **66** 736–743.

HUDSON, D. J. (1966). Fitting segmented curves whose join points have to be estimated. *J. Amer. Statist. Assoc.* **61** 1097–1129. MR0210243

JACK, C. R., PETERSEN, R. C., O'BRIEN, P. C. and TANGALOS, E. G. (1992). MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease. *Neurology* **42** 183–188.

JACK, C. R., PETERSEN, R. C., XU, Y. C., WARING, S. C., O'BRIEN, P. C., TANGALOS, E. G., SMITH, G. E., IVNIK, R. J. and KOKMEN, E. (1997). Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology* **49** 786–794.

KANTARCI, K. K. and JACK, C. R. (2003). Neuroimaging in Alzheimer disease: An evidence-based review. *Neuroimaging Clin. N. Am.* **13** 197–209.

LARRIEU, S., LETENNEUR, L., ORGOGOZO, J. M., FABRIGOULE, C., AMIEVA, H., LE CARRET, N., BARBERGER-GATEAU, P. and DARTIGUES, J. F. (2002). Incidence and outcome of mild cognitive impairment in a population-based prospective cohort. *Neurology* **59** 1594–1599.

LI, Y., QIAN, L. and ZHANG, W. (2013). Estimation in a change-point hazard regression model with long-term survivors. *Statist. Probab. Lett.* **83** 1683–1691. MR3062282

MA, J., MILLER, M. I. and YOUNES, L. (2010). A Bayesian generative model for surface template estimation. *Int. J. Biomed. Imaging* **2010** 974957.

MILLER, M. I., TROUVÉ, A. and YOUNES, L. (2015). Hamiltonian systems in computational anatomy: 100 years since D'Arcy Thompson. *Annu. Rev. Biomed. Eng.* **17**.

MILLER, M. I., YOUNES, L. and TROUVÉ, A. (2014). Diffeomorphometry and geodesic positioning systems for human anatomy. *Technology* **2** 36–43.

MILLER, M. I., YOUNES, L., RATNANATHER, J. T., BROWN, T., TRINH, H., LEE, D. S., TWARD, D., MAHON, P. B., MORI, S. and ALBERT, M. (2015). Amygdalar atrophy in symptomatic Alzheimer's disease based on diffeomorphometry: The BIOCARD cohort. *Neurobiol. Aging* **36** S3–S10. Supplement 1: Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders (NIBAD).

MUELLER, S. G., WEINER, M. W., THAL, L. J., PETERSEN, R. C., JACK, C. R., JAGUST, W., TROJANOWSKI, J. Q., TOGA, A. W. and BECKETT, L. (2005). Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's Dement.* **1** 55–66.

NGUYEN, H. T., ROGERS, G. S. and WALKER, E. A. (1984). Estimation in change-point hazard rate models. *Biometrika* **71** 299–304. MR767158

PIERSON, R., JOHNSON, H., HARRIS, G., KEEFE, H., PAULSEN, J. S., ANDREASEN, N. C. and MAGNOTTA, V. A. (2011). Fully automated analysis using BRAINS: AutoWorkup. *NeuroImage* **54** 328–336.

PONS, O. (2003). Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Ann. Statist.* **31** 442–463. Dedicated to the memory of Herbert E. Robbins. MR1983537

POULIN, S. P., DAUTOFF, R., MORRIS, J. C., BARRETT, L. F., DICKERSON, B. C., ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE et al. (2011). Amygdala atrophy is

prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry Research. Neuroimaging* **194** 7–13.

PRICE, J. L. (2003). Comparative aspects of amygdala connectivity. *Ann. N.Y. Acad. Sci.* **985** 50–58.

QUANDT, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *J. Amer. Statist. Assoc.* **53** 873–880. MR0100314

REUTER, M. (2010). Hierarchical shape segmentation and registration via topological features of Laplace–Beltrami eigenfunctions. *Int. J. Comput. Vis.* **89** 287–308.

RUSINEK, H., DE SANTI, S., FRID, D., TSUI, W.-H., TARSHISH, C. Y., CONVIT, A. and DE LEON, M. J. (2003). Regional brain atrophy rate predicts future cognitive decline: 6-year longitudinal MR imaging study of normal aging. *Radiology* **229** 691–696.

SCOTT, S. A., DEKOSKY, S. T. and SCHEFF, S. W. (1991). Volumetric atrophy of the amygdala in Alzheimer's disease: Quantitative serial reconstruction. *Neurology* **41** 351–356.

SCOTT, S. A., SPARKS, D. L., SCHEFF, S. W., DEKOSKY, S. T. and KNOX, C. A. (1992). Amygdala cell loss and atrophy in Alzheimer's disease. *Ann. Neurol.* **32** 555–563.

SHELINE, Y. I., GADO, M. H. and PRICE, J. L. (1998). Amygdala core nuclei volumes are decreased in recurrent major depression. *NeuroReport* **9** 2023–2028.

SPRENT, P. (1961). Some hypotheses concerning two phase regression lines. *Biometrics* **17** 634–645.

TANG, X., OISHI, K., FARIA, A. V., HILLIS, A. E., ALBERT, M. S., MORI, S. and MILLER, M. I. (2013). Bayesian parameter estimation and segmentation in the multi-atlas random orbit model. *PLoS ONE* **8** e65591.

TANG, X., HOLLAND, D., DALE, A. M., YOUNES, L., MILLER, M. I. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2014). Shape abnormalities of subcortical and ventricular structures in mild cognitive impairment and Alzheimer's disease: Detecting, quantifying, and predicting. *Hum. Brain Mapp.* **35** 3701–3725.

TANG, X., HOLLAND, D., DALE, A. M., YOUNES, L., MILLER, M. I. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2015). The diffeomorphometry of regional shape change rates and its relevance to cognitive deterioration in mild cognitive impairment and Alzheimer's disease. *Hum. Brain Mapp.* **36** 2093–2117.

THAMBISETTY, M., SIMMONS, A., VELAYUDHAN, L., HYE, A., CAMPBELL, J., ZHANG, Y., WAHLUND, L.-O., WESTMAN, E., KINSEY, A., GÜNTERT, A. et al. (2010). Association of plasma clusterin concentration with severity, pathology, and progression in Alzheimer disease. *Arch. Gen. Psychiatry* **67** 739–748.

TSUCHIYA, K. and KOSAKA, K. (1990). Neuropathological study of the amygdala in presenile Alzheimer's disease. *J. Neurol. Sci.* **100** 165–173.

WU, C. Q., ZHAO, L. C. and WU, Y. H. (2003). Estimation in change-point hazard function models. *Statist. Probab. Lett.* **63** 41–48. MR1973402

YOUNES, L. (2010). *Shapes and Diffeomorphisms. Applied Mathematical Sciences* **171**. Springer, Berlin. MR2656312

YOUNES, L., ALBERT, M., MILLER, M. I., BIOCARD RESEARCH TEAM et al. (2014). Inferring changepoint times of medial temporal lobe morphometric change in preclinical Alzheimer's disease. *NeuroImage Clin.* **5** 178–187.

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1763–1786
DOI: 10.1214/17-AOAS1057

DOUBLY ROBUST ESTIMATION OF OPTIMAL TREATMENT REGIMES FOR SURVIVAL DATA—WITH APPLICATION TO AN HIV/AIDS STUDY[1]

BY RUNCHAO JIANG*, WENBIN LU*,[2], RUI SONG*,[2], MICHAEL G. HUDGENS†,[3] AND SONIA NAPRVAVNIK†

*North Carolina State University** *and University of North Carolina at Chapel Hill*†

In many biomedical settings, assigning every patient the same treatment may not be optimal due to patient heterogeneity. Individualized treatment regimes have the potential to dramatically improve clinical outcomes. When the primary outcome is censored survival time, a main interest is to find optimal treatment regimes that maximize the survival probability of patients. Since the survival curve is a function of time, it is important to balance short-term and long-term benefit when assigning treatments. In this paper, we propose a doubly robust approach to estimate optimal treatment regimes that optimize a user specified function of the survival curve, including the restricted mean survival time and the median survival time. The empirical and asymptotic properties of the proposed method are investigated. The proposed method is applied to a data set from an ongoing HIV/AIDS clinical observational study conducted by the University of North Carolina (UNC) Center of AIDS Research (CFAR), and shows the proposed methods significantly improve the restricted mean time of the initial treatment duration. Finally, the proposed methods are extended to multi-stage studies.

REFERENCES

BAI, X., TSIATIS, A. A. and O'BRIEN, S. M. (2013). Doubly-robust estimators of treatment-specific survival distributions in observational studies with stratified sampling. *Biometrics* **69** 830–839. MR3146779

CHEN, P.-Y. and TSIATIS, A. A. (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics* **57** 1030–1038. MR1950418

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

DOMBROWSKI, J. C., KITAHATA, M. M., ROMPAEY, S. E. V., CRANE, H. M., MUGAVERO, M. J., ERON, J. J., BOSWELL, S. L., RODRIGUEZ, B., MATHEWS, W. C., MARTIN, J. N., MOORE, R. D. and GOLDEN, M. R. (2013). High levels of antiretroviral use and viral suppression among persons in HIV care in the United States, 2010. *J. Acquir. Immune Defic. Syndr.* **63** 299–306.

GILL, R. D., KEIDING, N. and ANDERSEN, P. K. (1997). *Statistical Models Based on Counting Processes*. Springer, New York.

GOLDBERG, Y. and KOSOROK, M. R. (2012). Q-learning with censored data. *Ann. Statist.* **40** 529–560. MR3014316

Key words and phrases. Doubly robust estimation, median survival time, optimal treatment regimen, restricted mean survival time.

GUNTHARD, H. F., ABERG, J. A., ERON, J. J., HOY, J. F., TELENTI, A., BENSON, C. A., BURGER, D. M., CAHN, P., GALLANT, J. E., GLESBY, M. J. REISS, P. SAAG, M. S. THOMAS, D. L. JACOBSEN, D. M. and VOLBERDING, P. A. (2014). Antiretroviral treatment of adult HIV infection: 2014 recommendations of the International Antiviral Society-USA panel. *JAMA J. Am. Med. Assoc.* **312** 410–425.

HOWE, C. J., COLE, S. R., NAPRAVNIK, S. and ERON, J. J.JR. (2010). Enrollment, retention, and visit attendance in the University of North Carolina Center for AIDS Research Clinical Cohort, 2001–2007. *AIDS Res. Hum. Retrovir.* **26** 875–881.

IRWIN, J. O. (1949). The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *J. Hyg.* **47** 188–189.

JIANG, R., LU, W., SONG, R., and DAVIDIAN, M. (2016). On estimation of optimal treatment regimes for maximizing t-year survival probability. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* To appear.

JIANG, R., LU, W., SONG, R., HUDGENS, M. G. and NAPRVAVNIK, S. (2017). Supplement to "Doubly robust estimation of optimal treatment regimes for survival data—with application to an HIV/AIDS study." DOI:10.1214/17-AOAS1057SUPP.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481. MR0093867

MEBANE, W. R. JR. and SEKHON, J. S. (2011). Genetic optimization using derivatives: The rgenoud package for R. *J. Stat. Softw.* **42** 1–26.

MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 331–366. MR1983752

MURPHY, S. A. (2005). A generalization error for Q-learning. *J. Mach. Learn. Res.* **6** 1073–1097. MR2249849

PANEL ON ANTIRETROVIRAL GUIDELINES FOR ADULTS AND ADOLESCENTS (2016). Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents, Department of Health and Human Services. Available at: https://aidsinfo.nih.gov/contentfiles/lvguidelines/adultandadolescentgl.pdf.

ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics. Lect. Notes Stat.* **179** 189–326. Springer, New York. MR2129402

TIAN, L., ALIZADEH, A. A., GENTLES, A. J. and TIBSHIRANI, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *J. Amer. Statist. Assoc.* **109** 1517–1532. MR3293607

WATKINS, C. J. C. H. and DAYAN, P. (1992). Q-learning. *Mach. Learn.* **8** 279–292.

WILLIG, J. H., ABROMS, S., WESTFALL, A. O., ROUTMAN, J., ADUSUMILLI, S., VARSHNEY, M., ALLISON, J., CHATHAM, A., RAPER, J. L., KASLOW, R. A., SAAG, M. S. and MUGAVERO, M. J. (2008). Increased regimen durability in the era of once daily fixed-dose combination antiretroviral therapy. *AIDS* **22** 1951–1960.

ZHAO, Y., KOSOROK, M. R. and ZENG, D. (2009). Reinforcement learning design for cancer clinical trials. *Stat. Med.* **28** 3294–3315. MR2750277

ZHAO, Y. Q., ZENG, D., LABER, E. B., SONG, R., YUAN, M. and KOSOROK, M. R. (2015). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika* **102** 151–168. MR3335102

ZUCKER, D. M. (1998). Restricted mean life with covariates: Modification and extension of a useful survival analysis method. *J. Amer. Statist. Assoc.* **93** 702–709. MR1631365

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1787–1809
DOI: 10.1214/17-AOAS1059

DYNAMIC PREDICTION FOR MULTIPLE REPEATED MEASURES AND EVENT TIME DATA: AN APPLICATION TO PARKINSON'S DISEASE

BY JUE WANG*, SHENG LUO*,[1] AND LIANG LI†

*University of Texas Health Science Center at Houston** *and*
University of Texas MD Anderson Cancer Center†

In many clinical trials studying neurodegenerative diseases such as Parkinson's disease (PD), multiple longitudinal outcomes are collected to fully explore the multidimensional impairment caused by this disease. If the outcomes deteriorate rapidly, patients may reach a level of functional disability sufficient to initiate levodopa therapy for ameliorating disease symptoms. An accurate prediction of the time to functional disability is helpful for clinicians to monitor patients' disease progression and make informative medical decisions. In this article, we first propose a joint model that consists of a semiparametric multilevel latent trait model (MLLTM) for the multiple longitudinal outcomes, and a survival model for event time. The two submodels are linked together by an underlying latent variable. We develop a Bayesian approach for parameter estimation and a dynamic prediction framework for predicting target patients' future outcome trajectories and risk of a survival event, based on their multivariate longitudinal measurements. Our proposed model is evaluated by simulation studies and is applied to the DATATOP study, a motivating clinical trial assessing the effect of deprenyl among patients with early PD.

REFERENCES

BLANCHE, P., PROUST-LIMA, C., LOUBÈRE, L., BERR, C., DARTIGUES, J.-F. and JACQMIN-GADDA, H. (2015). Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics* **71** 102–113. MR3335354

BROOKS, D. J. (2008). Optimizing levodopa therapy for Parkinson's disease with levodopa/carbidopa/entacapone: Implications from a clinical and patient perspective. *Neuropsychiatric Disease and Treatment* **4** 39–47.

BROWN, E. R. and IBRAHIM, J. G. (2003). Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59** 686–693. MR2004274

CHI, Y.-Y. and IBRAHIM, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* **62** 432–445. MR2227491

CRAINICEANU, C., RUPPERT, D. and WAND, M. P. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *J. Stat. Softw.* **14** 1–24.

DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998). Automatic Bayesian curve fitting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 333–350. MR1616029

DIMATTEO, I., GENOVESE, C. R. and KASS, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88** 1055–1071. MR1872219

Key words and phrases. Area under the ROC curve, clinical trial, failure time, latent trait model.

DUANE, S., KENNEDY, A. D., PENDLETON, B. J. and ROWETH, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195** 216–222.

DUNSON, D. B. (2007). Bayesian methods for latent trait modelling of longitudinal data. *Stat. Methods Med. Res.* **16** 399–415. MR2405478

ELASHOFF, R. M., LI, G. and LI, N. (2007). An approach to joint analysis of longitudinal measurements and competing risks failure time data. *Stat. Med.* **26** 2813–2835. MR2370939

ESCOBAR, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89** 268–277. MR1266299

FOX, J.-P. (2005). Multilevel IRT using dichotomous and polytomous response data. *Br. J. Math. Stat. Psychol.* **58** 145–172. MR2196136

FRIEDMAN, J. H. and SILVERMAN, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics* **31** 3–39. MR0997668

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. CRC Press, Boca Raton, FL. MR3235677

GILKS, W. R., BEST, N. G. and TAN, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Appl. Stat.* **44** 455–472.

HARRELL, F. (2015). *Regression Modeling Strategies*: *With Applications to Linear Models*, *Logistic and Ordinal Regression*, *and Survival Analysis*. Springer, Berlin.

HE, B. and LUO, S. (2016). Joint modeling of multivariate longitudinal measurements and survival data with applications to Parkinson's disease. *Stat. Methods Med. Res.* **25** 1346–1358. MR3541101

HENDERSON, R., DIGGLE, P. and DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1** 465–480.

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

IBRAHIM, J. G., CHU, H. and CHEN, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *J. Clin. Oncol.* **28** 2796–2801.

JACQMIN-GADDA, H., SIBILLOT, S., PROUST, C., MOLINA, J.-M. and THIÉBAUT, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Comput. Statist. Data Anal.* **51** 5142–5154. MR2370713

LAMBERT, P. and VANDENHENDE, F. (2002). A copula-based model for multivariate non-normal longitudinal data: Analysis of a dose titration safety study on a new antidepressant. *Stat. Med.* **21** 3197–3217.

LEE, S.-Y. and SONG, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivar. Behav. Res.* **39** 653–686.

LI, L., GREENE, T. and HU, B. (2016). A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Stat. Methods Med. Res.* Published online on Nov. 28, 2016, DOI:10.1177/0962280216680239.

LIU, L. and HUANG, X. (2009). Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 65–81. MR2662234

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.

LUO, S. and WANG, J. (2014). Bayesian hierarchical model for multiple repeated measures and survival data: An application to Parkinson's disease. *Stat. Med.* **33** 4279–4291. MR3267410

MCCULLOCH, C. E. and NEUHAUS, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statist. Sci.* **26** 388–402. MR2917962

MOLENBERGHS, G. and VERBEKE, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York. MR2171048

O'BRIEN, L. M. and FITZMAURICE, G. M. (2004). Analysis of longitudinal multiple-source binary data using generalized estimating equations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **53** 177–193. MR2043767

PROUST-LIMA, C., AMIEVA, H. and JACQMIN-GADDA, H. (2013). Analysis of multivariate mixed longitudinal data: A flexible latent process approach. *Br. J. Math. Stat. Psychol.* **66** 470–486. MR3120963

PROUST-LIMA, C., DARTIGUES, J.-F. and JACQMIN-GADDA, H. (2016). Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: A latent process and latent class approach. *Stat. Med.* **35** 382–398. MR3455508

PROUST-LIMA, C., SÉNE, M., TAYLOR, J. M. G. and JACQMIN-GADDA, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Stat. Methods Med. Res.* **23** 74–90. MR3190688

RIZOPOULOS, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67** 819–829. MR2829256

RIZOPOULOS, D., VERBEKE, G. and MOLENBERGHS, G. (2008). Shared parameter models under random effects misspecification. *Biometrika* **95** 63–74. MR2409715

RIZOPOULOS, D., MURAWSKA, M., ANDRINOPOULOU, E.-R., MOLENBERGHS, G., TAKKENBERG, J. J. and LESAFFRE, E. (2013). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. arXiv preprint arXiv:1306.6479.

RIZOPOULOS, D., HATFIELD, L. A., CARLIN, B. P. and TAKKENBERG, J. J. M. (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *J. Amer. Statist. Assoc.* **109** 1385–1397. MR3293598

RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.* **11** 735–757. MR1944261

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

SÈNE, M., TAYLOR, J. M. G., DIGNAM, J. J., JACQMIN-GADDA, H. and PROUST-LIMA, C. (2016). Individualized dynamic prediction of prostate cancer recurrence with and without the initiation of a second treatment: Development and validation. *Stat. Methods Med. Res.* **25** 2972–2991. MR3572894

SHOULSON, I. (1998). DATATOP: A decade of neuroprotective inquiry. Parkinson study group. Deprenyl and Tocopherol antioxidative therapy of Parkinsonism. *Ann. Neurol.* **44** S160–S166.

STAN DEVELOPMENT TEAM (2016). Stan modeling language users guide and reference manual, version 2.14.0.

STONE, C. J., HANSEN, M. H., KOOPERBERG, C. and TRUONG, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist.* **25** 1371–1470. MR1463561

SUN, J., PARK, D.-H., SUN, L. and ZHAO, X. (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *J. Amer. Statist. Assoc.* **100** 882–889. MR2201016

TAYLOR, J. M. G., PARK, Y., ANKERST, D. P., PROUST-LIMA, C., WILLIAMS, S., KESTIN, L., BAE, K., PICKLES, T. and SANDLER, H. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics* **69** 206–213. MR3058067

TSENG, Y.-K., HSIEH, F. and WANG, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92** 587–603. MR2202648

TSIATIS, A. A. and DAVIDIAN, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statist. Sinica* **14** 809–834. MR2087974

VAN HOUWELINGEN, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scand. J. Stat.* **34** 70–85. MR2325243

VERBEKE, G., FIEUWS, S., MOLENBERGHS, G. and DAVIDIAN, M. (2014). The analysis of multivariate longitudinal data: A review. *Stat. Methods Med. Res.* **23** 42–59. MR3190686

VONESH, E. F., GREENE, T. and SCHLUCHTER, M. D. (2006). Shared parameter models for the joint analysis of longitudinal data and event times. *Stat. Med.* **25** 143–163. MR2222079

WAND, M. P. (2000). A comparison of regression spline smoothing procedures. *Comput. Statist.* **15** 443–462. MR1818029

WANG, J. and LUO, S. (2017). Multidimensional latent trait linear mixed model: An application in clinical studies with multivariate longitudinal outcomes. *Stat. Med.* **36** 3244–3256, DOI:10.1002/sim.7347.

WANG, J., LUO, S. and LI, L. (2017). Supplement to "Dynamic prediction for multiple repeated measures and event time data: An application to Parkinson's disease." DOI:10.1214/17-AOAS1059SUPP.

WULFSOHN, M. S. and TSIATIS, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53** 330–339. MR1450186

XU, J. and ZEGER, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **50** 375–387. MR1856332

YANG, L., YU, M. and GAO, S. (2016). Prediction of coronary artery disease risk based on multiple longitudinal biomarkers. *Stat. Med.* **35** 1299–1314. MR3512813

The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1810–1831
DOI: 10.1214/17-AOAS1062

TESTING HIGH-DIMENSIONAL COVARIANCE MATRICES, WITH APPLICATION TO DETECTING SCHIZOPHRENIA RISK GENES

BY LINGXUE ZHU[*], JING LEI[*,1], BERNIE DEVLIN[†,2] AND KATHRYN ROEDER[*,2,3]

Carnegie Mellon University[*] and University of Pittsburgh[†]

Scientists routinely compare gene expression levels in cases versus controls in part to determine genes associated with a disease. Similarly, detecting case-control differences in co-expression among genes can be critical to understanding complex human diseases; however, statistical methods have been limited by the high-dimensional nature of this problem. In this paper, we construct a sparse-Leading-Eigenvalue-Driven (sLED) test for comparing two high-dimensional covariance matrices. By focusing on the spectrum of the differential matrix, sLED provides a novel perspective that accommodates what we assume to be common, namely sparse and weak signals in gene expression data, and it is closely related with sparse principal component analysis. We prove that sLED achieves full power asymptotically under mild assumptions, and simulation studies verify that it outperforms other existing procedures under many biologically plausible scenarios. Applying sLED to the largest gene-expression dataset obtained from post-mortem brain tissue from Schizophrenia patients and controls, we provide a novel list of genes implicated in Schizophrenia and reveal intriguing patterns in gene co-expression change for Schizophrenia subjects. We also illustrate that sLED can be generalized to compare other gene-gene "relationship" matrices that are of practical interest, such as the weighted adjacency matrices.

REFERENCES

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York. MR0091588

BAI, Z., JIANG, D., YAO, J.-F. and ZHENG, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.* **37** 3822–3840. MR2572444

BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849

CAI, T., LIU, W. and XIA, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.* **108** 265–277. MR3174618

CAI, T. T. and ZHANG, A. (2016). Inference for high-dimensional differential correlation matrices. *J. Multivariate Anal.* **143** 107–126. MR3431422

CHANG, J., ZHOU, W., ZHOU, W.-X. and WANG, L. (2016). Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. Preprint. Available at arXiv:1505.04493v3.

Key words and phrases. Permutation test, high-dimensional data, covariance matrix, sparse principal component analysis.

CHEN, E. Y., TAN, C. M., KOU, Y., DUAN, Q., WANG, Z., MEIRELLES, G. V., CLARK, N. R. and MA'AYAN, A. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* **14** 128.

D'ASPREMONT, A., EL GHAOUI, L., JORDAN, M. I. and LANCKRIET, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* **49** 434–448. MR2353806

DINIZ, L. P., ALMEIDA, J. C., TORTELLI, V., VARGAS LOPES, C., SETTI-PERDIGÃO, P., STIPURSKY, J., KAHN, S. A., ROMÃO, L. F., DE MIRANDA, J., ALVES-LEON, S. V., DE SOUZA, J. M., CASTRO, N. G., PANIZZUTTI, R. and GOMES, F. C. A. (2012). Astrocyte-induced synaptogenesis is mediated by transforming growth factor β signaling through modulation of D-serine levels in cerebral cortex neurons. *J. Biol. Chem.* **287** 41432–41445.

FROMER, M., ROUSSOS, P., SIEBERTS, S. K., JOHNSON, J. S., KAVANAGH, D. H., PERUMAL, T. M., RUDERFER, D. M., OH, E. C., TOPOL, A., SHAH, H. R., KLEI, L. L., KRAMER, R., PINTO, D., GÜMÜŞ, Z. H., CICEK, A. E., DANG, K. K., BROWNE, A., LU, C., XIE, L., READHEAD, B., STAHL, E. A., XIAO, J., PARVIZI, M., HAMAMSY, T., FULLARD, J. F., WANG, Y.-C., MAHAJAN, M. C., DERRY, J. M. J., DUDLEY, J. T., HEMBY, S. E., LOGSDON, B. A., TALBOT, K., RAJ, T., BENNETT, D. A., DE JAGER, P. L., ZHU, J., ZHANG, B., SULLIVAN, P. F., CHESS, A., PURCELL, S. M., SHINOBU, L. A., MANGRAVITE, L. M., TOYOSHIBA, H., GUR, R. E., HAHN, C.-G., LEWIS, D. A., HAROUTUNIAN, V., PETERS, M. A., LIPSKA, B. K., BUXBAUM, J. D., SCHADT, E. E., HIRAI, K., ROEDER, K., BRENNAND, K. J., KATSANIS, N., DOMENICI, E., DEVLIN, B. and SKLAR, P. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19** 1442–1453.

GU, X., LI, A., LIU, S., LIN, L., XU, S., ZHANG, P., LI, S., LI, X., TIAN, B., ZHU, X. and WANG, X. (2015). MicroRNA124 regulated neurite elongation by targeting OSBP. *Mol. Neurobiol.* **53** 6388–6396.

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

JOLLIFFE, I. T., TRENDAFILOV, N. T. and UDDIN, M. (2003). A modified principal component technique based on the LASSO. *J. Comput. Graph. Statist.* **12** 531–547. MR2002634

LI, J. and CHEN, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *Ann. Statist.* **40** 908–940. MR2985938

MCGRATH, J., SAHA, S., CHANT, D. and WELHAM, J. (2008). Schizophrenia: A concise overview of incidence, prevalence, and mortality. *Epidemiol. Rev.* **30** 67–76.

OWEN, M. J., SAWA, A. and MORTENSEN, P. B. (2016). Schizophrenia. *Lancet* **388** 86–97.

OYMAK, S., JALALI, A., FAZEL, M., ELDAR, Y. C. and HASSIBI, B. (2015). Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Trans. Inform. Theory* **61** 2886–2908. MR3342310

INTERNATIONAL SCHIZOPHRENIA CONSORTIUM, PURCELL, S. M., WRAY, N. R., STONE, J. L., VISSCHER, P. M., O'DONOVAN, M. C., SULLIVAN, P. F. and SKLAR, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460** 748–752.

PURCELL, S. M., MORAN, J. L., FROMER, M., RUDERFER, D., SOLOVIEFF, N., ROUSSOS, P., O'DUSHLAINE, C., CHAMBERT, K., BERGEN, S. E., KÄHLER, A., DUNCAN, L., STAHL, E., GENOVESE, G., FERNÁNDEZ, E., COLLINS, M. O., KOMIYAMA, N. H., CHOUDHARY, J. S., MAGNUSSON, P. K. E., BANKS, E., SHAKIR, K., GARIMELLA, K., FENNELL, T., DEPRISTO, M., GRANT, S. G. N., HAGGARTY, S. J., GABRIEL, S., SCOLNICK, E. M., LANDER, E. S., HULTMAN, C. M., SULLIVAN, P. F., MCCARROLL, S. A. and SKLAR, P. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506** 185–90.

SCHIZOPHRENIA WORKING GROUP OF THE PSYCHIATRIC GENOMICS CONSORTIUM (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511** 421–427.

SCHOTT, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Comput. Statist. Data Anal.* **51** 6535–6542. MR2408613

SHEN, H. and HUANG, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.* **99** 1015–1034. MR2419336

SRIVASTAVA, M. S. and YANAGIHARA, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *J. Multivariate Anal.* **101** 1319–1329. MR2609494

SZÉKELY, G. J. and RIZZO, M. L. (2013). Energy statistics: A class of statistics based on distances. *J. Statist. Plann. Inference* **143** 1249–1272. MR3055745

VU, V. Q., CHO, J., LEI, J. and ROHE, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *Advances in Neural Information Processing Systems* **26** 2670–2678.

WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

WU, T.-L. and LI, P. (2015). Tests for high-dimensional covariance matrices using random matrix projection. Preprint. Available at arXiv:1511.01611.

YU, C.-Y., GUI, W., HE, H.-Y., WANG, X.-S., ZUO, J., HUANG, L., ZHOU, N., WANG, K. and WANG, Y. (2014). Neuronal and astroglial TGFβ-Smad3 signaling pathways differentially regulate dendrite growth and synaptogenesis. *Neuromolecular Med.* **16** 457–72.

ZHANG, B. and HORVATH, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 17, 45. MR2170433

ZHU, L., LEI, J., DEVLIN, B. and ROEDER, K. (2017). Supplement to "Testing high-dimensional covariance matrices, with application to detecting schizophrenia risk genes." DOI:10.1214/17-AOAS1062SUPP.

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527