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*

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1137–1156
DOI: 10.1214/15-AOAS880

FIBER DIRECTION ESTIMATION, SMOOTHING AND TRACKING IN DIFFUSION MRI[1,2]

BY RAYMOND K. W. WONG*, THOMAS C. M. LEE†,3, DEBASHIS PAUL†,4, JIE PENG†,5 AND ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

Iowa State University and University of California, Davis†*

Diffusion magnetic resonance imaging is an imaging technology designed to probe anatomical architectures of biological samples in an in vivo and noninvasive manner through measuring water diffusion. The contribution of this paper is threefold. First, it proposes a new method to identify and estimate multiple diffusion directions within a voxel through a new and identifiable parametrization of the widely used multi-tensor model. Unlike many existing methods, this method focuses on the estimation of diffusion directions rather than the diffusion tensors. Second, this paper proposes a novel direction smoothing method which greatly improves direction estimation in regions with crossing fibers. This smoothing method is shown to have excellent theoretical and empirical properties. Last, this paper develops a fiber tracking algorithm that can handle multiple directions within a voxel. The overall methodology is illustrated with simulated data and a data set collected for the study of Alzheimer's disease by the Alzheimer's Disease Neuroimaging Initiative (ADNI).

REFERENCES

ARSIGNY, V., FILLARD, P., PENNEC, X. and AYACHE, N. (2006). Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magn. Reson. Med.* **56** 411–421.

BAMMER, R., HOLDSWORTH, S. J., VELDHUIS, W. B. and SKARE, S. T. (2009). New methods in diffusion-weighted and diffusion tensor imaging. *Magn. Reson. Imaging Clin. N. Am.* **17** 175–204.

BASSER, P. J., PAJEVIC, S., PIERPAOLI, C., DUDA, J. and ALDROUBI, A. (2000). In vivo fiber tractography using DT-MRI data. *Magn. Reson. Med.* **44** 625–632.

BEAULIEU, C. (2002). The basis of anisotropic water diffusion in the nervous system—A technical review. *NMR Biomed.* **15** 435–455.

BEHRENS, T. E. J., WOOLRICH, M. W., JENKINSON, M., JOHANSEN-BERG, H., NUNES, R. G., CLARE, S., MATTHEWS, P. M., BRADY, J. M. and SMITH, S. M. (2003). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magn. Reson. Med.* **50** 1077–1088.

BEHRENS, T. E. J., BERG, H. J., JBABDI, S., RUSHWORTH, M. F. S. and WOOLRICH, M. W. (2007). Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *Neuroimage* **34** 144–155.

CARMICHAEL, O., CHEN, J., PAUL, D. and PENG, J. (2013). Diffusion tensor smoothing through weighted Karcher means. *Electron. J. Stat.* **7** 1913–1956. MR3084676

Key words and phrases. Diffusion tensor imaging, direction smoothing, multi-tensor model, fiber tracking, tractography.

CHANRAUD, S., ZAHR, N., SULLIVAN, E. V. and PFEFFERBAUM, A. (2010). MR diffusion tensor imaging: A window into white matter integrity of the working brain. *Neuropsychol. Rev.* **20** 209–225.

DESCOTEAUX, M., ANGELINO, E., FITZGIBBONS, S. and DERICHE, R. (2007). Regularized, fast, and robust analytical Q-ball imaging. *Magn. Reson. Med.* **58** 497–510.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. Chapman & Hall, London. MR1383587

FILLARD, P., PENNEC, X., ARSIGNY, V. and AYACHE, N. (2007). Clinical DT-MRI estimation, smoothing, and fiber tracking with log-Euclidean metrics. *Medical Imaging, IEEE Transactions on* **26** 1472–1482.

FLETCHER, P. T. and JOSHI, S. (2007). Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing* **87** 250–262.

FRIMAN, O., FARNEBACK, G. and WESTIN, C.-F. (2006). A Bayesian approach for stochastic white matter tractography. *Medical Imaging, IEEE Transactions on* **25** 965–978.

GUDBJARTSSON, H. and PATZ, S. (1995). The Rician distribution of noisy MRI data. *Magn. Reson. Med.* **34** 910–914.

HOSEY, T., WILLIAMS, G. and ANSORGE, R. (2005). Inference of multiple fiber orientations in high angular resolution diffusion imaging. *Magn. Reson. Med.* **54** 1480–1489.

KAUFMAN, L. and ROUSSEEUW, P. J. (1990). *Finding Groups in Data*: *An Introduction to Cluster Analysis*. Wiley, New York. MR1044997

KOCH, M. A., NORRIS, D. G. and HUND-GEORGIADIS, M. (2002). An investigation of functional and anatomical connectivity using magnetic resonance imaging. *Neuroimage* **16** 241–250.

MORI, S. (2007). *Introduction to Diffusion Tensor Imaging*. Elsevier, Amsterdam.

MORI, S. and VAN ZIJL, P. C. M. (2002). Fiber tracking: Principles and strategies—A technical review. *NMR Biomed.* **15** 468–480.

MORI, S., CRAIN, B. J., CHACKO, V. P. and VAN ZIJL, P. (1999). Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of Neurology* **45** 265–269.

MUKHERJEE, P., BERMAN, J. I., CHUNG, S. W., HESS, C. P. and HENRY, R. G. (2008). Diffusion tensor MR imaging and fiber tractography: Theoretic underpinnings. *AJNR Am. J. Neuroradiol.* **29** 632–641.

NIMSKY, C., GANSLANDT, O. and FAHLBUSCH, R. (2006). Implementation of fiber tract navigation. *Neurosurgery* **58** ONS–292–303; discussion ONS–303–4.

PARKER, G. J. M. and ALEXANDER, D. C. (2003). Probabilistic Monte Carlo based mapping of cerebral connections utilising whole-brain crossing fibre information. In *Information Processing in Medical Imaging* 684–695. Springer, Berlin.

PENNEC, X., FILLARD, P. and AYACHE, N. (2006). A Riemannian framework for tensor computing. *Int. J. Comput. Vis.* **66** 41–66.

ROUSSEEUW, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20** 53–65.

SCHERRER, B. and WARFIELD, S. K. (2010). Why multiple b-values are required for multi-tensor models. Evaluation with a constrained log-Euclidean model. In 2010 *IEEE International Symposium on Biomedical Imaging*: *From Nano to Macro* 1389–1392. IEEE, Rotterdam.

SCHWARTZMAN, A., DOUGHERTY, R. F. and TAYLOR, J. E. (2008). False discovery rate analysis of brain diffusion direction maps. *Ann. Appl. Stat.* **2** 153–175. MR2415598

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SPORNS, O. (2011). *Networks of the Brain*. MIT Press, Cambridge, MA.

TABELOW, K., VOSS, H. U. and POLZEHL, J. (2012). Modeling the orientation distribution function by mixtures of angular central Gaussian distributions. *J. Neurosci. Methods* **203** 200–211.

TOURNIER, J., CALAMANTE, F., GADIAN, D. G., CONNELLY, A. et al. (2004). Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. *NeuroImage* **23** 1176–1185.

TOURNIER, J., CALAMANTE, F., CONNELLY, A. et al. (2007). Robust determination of the fibre orientation distribution in diffusion MRI: Non-negativity constrained super-resolved spherical deconvolution. *NeuroImage* **35** 1459–1472.

TUCH, D. S. (2002). Diffusion MRI of complex tissue structure. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

TUCH, D. S. (2004). Q-ball imaging. *Magn. Reson. Med.* **52** 1358–1372.

TUCH, D. S., REESE, T. G., WIEGELL, M. R., MAKRIS, N., BELLIVEAU, J. W. and WEDEEN, V. J. (2002). High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. *Magn. Reson. Med.* **48** 577–582.

WEINSTEIN, D., KINDLMANN, G. and LUNDBERG, E. (1999). Tensorlines: Advection-diffusion based propagation through diffusion tensor fields. In *Proceedings of the Conference on Visualization* 249–253.

WIEGELL, M. R., LARSSON, H. B. and WEDEEN, V. J. (2000). Fiber crossing in human brain depicted with diffusion tensor MR imaging1. *Radiology* **217** 897–903.

WONG, R. K. W., LEE, T. C. M., PAUL, D. and PENG, J. (2016). Supplement to "Fiber direction estimation, smoothing and tracking in diffusion MRI." DOI:10.1214/15-AOAS880SUPP.

YUAN, Y., ZHU, H., LIN, W. and MARRON, J. S. (2012). Local polynomial regression for symmetric positive definite matrices. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 697–719. MR2965956

ZHU, H., ZHANG, H., IBRAHIM, J. G. and PETERSON, B. S. (2007). Statistical analysis of diffusion tensors in diffusion-weighted magnetic resonance imaging data. *J. Amer. Statist. Assoc.* **102** 1085–1102. MR2412530

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1157–1159
DOI: 10.1214/16-AOAS935
Main article DOI: 10.1214/15-AOAS880

DISCUSSION OF "FIBER DIRECTION ESTIMATION IN DIFFUSION MRI"

BY ARMIN SCHWARTZMAN

University of California, San Diego

REFERENCES

WATSON, G. S. (1965). Equatorial distributions on a sphere. *Biometrika* **52** 193–201. MR0207115

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1160–1161
DOI: 10.1214/16-AOAS936
Main article DOI: 10.1214/15-AOAS880

DISCUSSION OF "FIBER DIRECTION ESTIMATION IN DIFFUSION MRI"

BY NICOLE A. LAZAR

University of Georgia

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1162–1165
DOI: 10.1214/16-AOAS937
Main article DOI: 10.1214/15-AOAS880

DISCUSSION OF "FIBER DIRECTION ESTIMATION IN DIFFUSION MRI"

BY JIAN KANG[1] AND LEXIN LI[2]

University of Michiga and University of California, Berkeley

REFERENCES

FRIMAN, O., FARNEBÄCK, G. and WESTIN, C.-F. (2006). A Bayesian approach for stochastic white matter tractography. *Medical Imaging*, *IEEE Transactions on* **25** 965–978.

JONES, M. C., MARRON, J. S. and SHEATHER, S. J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.* **91** 401–407. MR1394097

RAYKAR, V. C. and DURAISWAMI, R. (2006). Fast optimal bandwidth selection for kernel density estimation. In *Proceedings of the Sixth SIAM International Conference on Data Mining* 524–528. SIAM, Philadelphia, PA. MR2337970

Key words and phrases. Diffusion tensor imaging, discussion.

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1166–1169
DOI: 10.1214/16-AOAS880R
Main article DOI: 10.1214/15-AOAS880

REJOINDER: "FIBER DIRECTION ESTIMATION, SMOOTHING AND TRACKING IN DIFFUSION MRI"

BY RAYMOND K. W. WONG*, THOMAS C. M. LEE†, DEBASHIS PAUL† AND JIE PENG†

*Iowa State University** *and University of California, Davis*†

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1170–1194
DOI: 10.1214/15-AOAS902

ASSESSING THE CAUSAL EFFECTS OF FINANCIAL AIDS TO FIRMS IN TUSCANY ALLOWING FOR INTERFERENCE

BY BRUNO ARPINO AND ALESSANDRA MATTEI[1]

Universitat Pompeu Fabra and University of Florence

We consider policy evaluations when the Stable Unit Treatment Value Assumption (SUTVA) is violated due to the presence of interference among units. We propose to explicitly model interference as a function of units' characteristics. Our approach is applied to the evaluation of a policy implemented in Tuscany (a region in Italy) on small handicraft firms. Results show that the benefits from the policy are reduced when treated firms are subject to high levels of interference. Moreover, the average causal effect is slightly underestimated when interference is ignored. We stress the importance of considering possible interference among units when evaluating and planning policy interventions.

REFERENCES

ALMUS, M. and CZARNITZKI, D. (2003). The effects of public R&D subsidies on firms' innovation activities: The case of Eastern Germany. *J. Bus. Econom. Statist.* **21** 226–236. MR1973746

ANSELIN, L. (2006). Spatial econometrics. In *Palgrave Handbook of Econometrics*: *Volume* 1, *Econometric Theory* (T. C. Mills and K. Patterson, eds.) 901–941. Palgrave Macmillan, Basingstoke, UK.

ARONOW, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociol. Methods Res.* **41** 3–16. MR3190698

BATTISTIN, E., GAVOSTO, A. and RETTORE, E. (2001). Why do subsidised firms survive longer? An evaluation of a program promoting youth entrepreneurship in Italy. In *Econometric Evaluation of Labour Market Policies* 153–181. Springer, Berlin.

BHATTACHARJEE, A. and JENSEN-BUTLER, C. (2013). Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics* **43** 617–634.

BIA, M. and MATTEI, A. (2012). Assessing the effect of the amount of financial aids to Piedmont firms using the generalized propensity score. *Stat. Methods Appl.* **21** 485–516. MR2992915

BLACKWELL, M., IACUS, S. M., KING, G. and PORRO, G. (2009). CEM: Coarsened exact matching in stata. *The Stata Journal* **9** 524–546.

BÖRNER, K., SANYAL, S. and VESPIGNANI, A. (2007). Network science. *Annual Review of Information Science and Technology* **41** 537–607.

BRAMOULLÉ, Y., DJEBBARI, H. and FORTIN, B. (2009). Identification of peer effects through social networks. *J. Econometrics* **150** 41–55. MR2525993

BROCK, W. A. and DURLAUF, S. N. (2001). Interactions-based models. In *Handbook of Econometrics* **5**. Elsevier, Amsterdam.

BRONZINI, R. and DE BLASIO, G. (2006). Evaluating the impact of investment incentives: The case of Italy's law 488/1992. *Journal of Urban Economics* **60** 327–349.

Key words and phrases. Interference, causal inference, policy evaluation, potential outcomes, Rubin Causal Model, SUTVA.

CORRADO, L. and FINGLETON, B. (2012). Where is the economics in spatial econometrics? *Journal of Regional Science* **52** 210–239.

CRÉPON, B., DUFLO, E., GURGAND, M., RATHELOT, R. and ZAMORA, P. (2013). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *The Quarterly Journal of Economics* **128** 531–580.

DE CASTRIS, M. and PELLEGRINI, G. (2012). Evaluation of spatial effects of capital subsidies in the South of Italy. *Regional Studies* **46** 525–538.

DEHEJIA, R. H. and WAHBA, S. (1999). Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *J. Amer. Statist. Assoc.* **94** 10053–1062.

DEI OTTATI, G. (1994). Cooperation and competition in the industrial district as an organization model. *European Planning Studies* **2** 463–483.

DOREIAN, P. (1996). When the data points are not independent. In *Developments in Data Analysis: Proceedings of the International Conference on Statistical Data Analysis and Data Collection, Bled, Slovenia, September* 19-21, 1994 (A. Ferligoj and A. Kramberger, eds.) 27–46. FDV, Fakulteta za družbene vede, Univerza v Ljubljani. Slovenia.

DUCH, N., MONTOLIO, D. and MEDIAVILLA, M. (2009). Evaluating the impact of public subsidies on a firm's performance: A two-stage quasi-experimental approach. *Investigaciones Regionales* **16** 143–165.

EBERHARDT, M., HELMERS, C. and STRAUSS, H. (2013). Do spillovers matter when estimating private returns to R&D? *The Review of Economics and Statistics* **95** 436–448.

GLEDITSCH, K. S., WARD, M. D. and KRISTIAN, S. (2007). An introduction to spatial regression models in the social sciences. Technical report, Duke Univ., Durham, NC.

GREINER, D. J. and RUBIN, D. B. (2011). Causal effects of perceived immutable characteristics. *The Review of Economics and Statistics* **93** 775–785.

HALLECK VEGA, S. and ELHORST, J. P. (2015). The SLX model. *Journal of Regional Science* **55** 339–363.

HARRIS, R., MOFFAT, J. and KRAVTSOVA, V. (2011). In search of "W". *Spatial Economic Analysis* **6** 249–270.

HIRANO, K. and IMBENS, W. G. (2004). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (A. Gelman and X. L. Meng, eds.) **226164** 73–84. Wiley, Hoboken, NJ.

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15** 199–236.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

HONG, G. and RAUDENBUSH, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J. Amer. Statist. Assoc.* **101** 901–910. MR2324091

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472

IACUS, S. M., KING, G. and PORRO, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis* **20** 1–24.

IMAI, K. and VAN DYK, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* **99** 854–866. MR2090918

IMBENS, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710. MR1789821

IMBENS, G. W. (2004). Nonparametric estimation of average treatment effetcs under exogeneity: A review. *The Review of Economics and Statistics* **86** 4–29.

KAO, E., TOULIS, R., AIROLDI, E. and RUBIN, B. D. (2012). Causal estimation of peer influence effects. In *NIPS 2012 Workshop 'Social Network and Social Media Analysis: Methods, Models and Applications'*. Stanford Network Analysis Project, Stanford University, Lake Tahoe, Nevada.

KLETTE, T., MØEN, J. and GRILICHES, Z. (2000). Do subsidies to commercial R&D reduce market failures? Microeconometric evaluation studies. *Research Policy* **29** 471–495.

LECHNER, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies* (M. Lechner and F. Pfeiffer, eds.). *ZEW Economic Studies* **13** 43–58. Physica-Verlag, Heidelberg.

MATTEI, A. and MAURO, V. (2007). Valutazione di politiche per le imprese artigiane. Research report. IRPET—Istituto Regionale Programmazione Economica della Toscana.

MEALLI, F. and RUBIN, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **102** 995–1000. MR3431570

PELLEGRINI, G. and CARLUCCI, C. (2003). Gli effetti della legge 488/92: Una valutazione dell'impatto occupazionale sulle imprese agevolate. *Rivista Italiana degli Economisti* **8** 267–286.

ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *J. Amer. Statist. Assoc*. **102** 191–200. MR2345537

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. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist*. **3** 33–38.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

RUBIN, D. B. (1980). Discussion of "Randomization analysis of experimental data: the Fisher randomization test" by D. Basu. *J. Amer. Statist. Assoc*. **75** 591–593.

RUBIN, D. B. (1990). Comment on J. Neyman and causal inference in experiments and observational studies: "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" [Ann. Agric. Sci. **1**0 (1923), 1–51]. *Statist. Sci*. **5** 472–480. MR1092987

SAMII, C. and ARONOW, P. (2013). Estimating average causal effects under general interference. Technical report. Available at http://arxiv.org/abs/1305.6156.

SOBEL, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Amer. Statist. Assoc*. **101** 1398–1407. MR2307573

STUART, E. A. (2007). Estimating causal effects using school-level data sets. *Educational Researcher* **36** 187–198.

TAKALO, T., TANAYAMA, T. and TOIVANEN, O. (2013). Estimating the benefits of targeted R&D subsidies. *The Review of Economics and Statistics* **95** 255–272.

TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res*. **21** 55–75. MR2867538

VERBITSKY, N. and RAUDENBUSH, S. W. (2004). Causal inference in spatial settings. In *Preceedings of the Social Statistics Section* 2369–2374. Amer. Statist. Assoc., Alexandria, VA.

WOOLDRIDGE, J. M. and IMBENS, G. W. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* **47** 5–86.

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1195–1216
DOI: 10.1214/16-AOAS906

DECONVOLUTION OF BASE PAIR LEVEL RNA-SEQ READ COUNTS FOR QUANTIFICATION OF TRANSCRIPT EXPRESSION LEVELS[1]

BY HAN WU* AND YU ZHU*,†

*Purdue University** *and Tsinghua University*†

RNA-Seq has emerged as the method of choice for profiling the transcriptomes of organisms. In particular, it aims to quantify the expression levels of transcripts using short nucleotide sequences or short reads generated from RNA-Seq experiments. In real experiments, the label of the transcript, from which each short read is generated, is missing, and short reads are mapped to the genome rather than the transcriptome. Therefore, the quantification of transcript expression levels is an indirect statistical inference problem.

In this article, we propose to use individual exonic base pairs as observation units and, further, to model nonzero as well as zero counts at all base pairs at both the transcript and gene levels. At the transcript level, two-component Poisson mixture distributions are postulated, which gives rise to the Convolution of Poisson mixture (CPM) distribution model at the gene level. The maximum likelihood estimation method equipped with the EM algorithm is used to estimate model parameters and quantify transcript expression levels. We refer to the proposed method as CPM-Seq. Both simulation studies and real data demonstrate the effectiveness of CPM-Seq, showing that CPM-Seq produces more accurate and consistent quantification results than Cufflinks.

REFERENCES

AU, K. F., JIANG, H., LIN, L., XING, Y. and WONG, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res*. **38** 4570–4578.

GRIEBEL, T., ZACHER, B., RIBECA, P., RAINERI, E., LACROIX, V., GUIGÓ, R. and SAMMETH, M. (2012). Modelling and simulating generic RNA-seq experiments with the flux simulator. *Nucleic Acids Res*. **40** 10073–10083.

HU, M., ZHU, Y., TAYLOR, J. M. G., LIU, J. S. and QIN, Z. S. (2012). Using Poisson mixed-effects model to quantify transcript-level gene expression in RNA-seq. *Bioinformatics* **28** 63–68.

KIM, H., BI, Y., PAL, S., GUPTA, R. and DAVULURI, R. V. (2011). IsoformEx: Isoform level gene expression estimation using weighted non-negative least squares from mRNA-seq data. *BMC Bioinformatics* **12** 305.

LANGMEAD, B., TRAPNELL, C., POP, M. and SALZBERG, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. **10** R25.

LI, B. and DEWEY, C. N. (2011). RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12** 323.

LI, W., FENG, J. and JIANG, T. (2011). IsoLasso: A LASSO regression approach to RNA-Seq based transcriptome assembly. *J. Comput. Biol*. **18** 1693–1707. MR2860071

Key words and phrases. RNA-Seq, transcriptome profiling, finite Poisson mixture model, convolution.

LI, J. J., JIANG, C.-R., BROWN, J. B., HUANG, H. and BICKEL, P. J. (2011). Sparse linear modeling of next-generation mRNA sequencing (RNA-seq) data for isoform discovery and abundance estimation. *Proc. Natl. Acad. Sci. USA* **108** 19867–19872.

MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L. and WOLD, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* **5** 621–628.

SALZMAN, J., JIANG, H. and WONG, W. H. (2011). Statistical modeling of RNA-Seq data. *Statist. Sci.* **26** 62–83. MR2849910

SRIVASTAVA, S. and CHEN, L. (2010). A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res.* **38** e170.

TRAPNELL, C., PACHTER, L. and SALZBERG, S. L. (2009). TopHat: Discovering splice junctions with RNA-seq. *Bioinformatics* **25** 1105–1111.

TRAPNELL, C., WILLIAMS, B. A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M. J., SALZBERG, S. L., WOLD, B. J. and PACHTER, L. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28** 511–515.

VARIN, C., REID, N. and FIRTH, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42. MR2796852

WANG, E. T., SANDBERG, R., LUO, S., KHREBTUKOVA, I., ZHANG, L., MAYR, C., KINGSMORE, S. F., SCHROTH, G. P. and BURGE, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* **456** 470–476.

WU, H., QIN, Z. S. and ZHU, Y. (2012). PM-Seq: Using finite Poisson mixture models for RNA-seq data analysis and transcript expression level quantification. *Statistics in Biosciences* **5** 71–87.

WU, H. and ZHU, Y. (2016). Supplement to "Deconvolution of base pair level RNA-seq read counts for quantification of transcript expression levels." DOI:10.1214/16-AOAS906SUPP.

ZAHARIA, M., CHOWDHURY, M., FRANKLIN, M. J., SHENKER, S. and STOICA, I. (2010). Spark: Cluster computing with working sets. In *HotCloud'*10 *Proceedings of the* 2*Nd USENIX Conference on Hot Topics in Cloud Computing* 10–10. USENIX Association, Berkeley, CA.

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1217–1244
DOI: 10.1214/16-AOAS907

GENE-PROXIMITY MODELS FOR GENOME-WIDE ASSOCIATION STUDIES[1]

BY IAN JOHNSTON[*], TIMOTHY HANCOCK[†], HIROSHI MAMITSUKA[†,2] AND LUIS CARVALHO[*,3]

Boston University[*] and Kyoto University[†]

Motivated by the important problem of detecting association between genetic markers and binary traits in genome-wide association studies, we present a novel Bayesian model that establishes a hierarchy between markers and genes by defining weights according to gene lengths and distances from genes to markers. The proposed hierarchical model uses these weights to define unique prior probabilities of association for markers based on their proximities to genes that are believed to be relevant to the trait of interest. We use an expectation-maximization algorithm in a filtering step to first reduce the dimensionality of the data and then sample from the posterior distribution of the model parameters to estimate posterior probabilities of association for the markers. We offer practical and meaningful guidelines for the selection of the model tuning parameters and propose a pipeline that exploits a singular value decomposition on the raw data to make our model run efficiently on large data sets. We demonstrate the performance of the model in simulation studies and conclude by discussing the results of a case study using a real-world data set provided by the Wellcome Trust Case Control Consortium.

REFERENCES

1000 GENOMES PROJECT CONSORTIUM et al. (2012). An integrated map of genetic variation from 1092 human genomes. *Nature* **491** 56–65.

AL-MUBAID, H. and SINGH, R. K. (2010). A text-mining technique for extracting gene-disease associations from the biomedical literature. *International Journal of Bioinformatics Research and Applications* **6** 270–286.

BALDING, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* **7** 781–791.

BANSAL, V., LIBIGER, O., TORKAMANI, A. and SCHORK, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* **11** 773–785.

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York. MR0804611

BURTON, P. R., CLAYTON, D. G., CARDON, L. R., CRADDOCK, N., DELOUKAS, P., DUNCANSON, A., KWIATKOWSKI, D. P., MCCARTHY, M. I., OUWEHAND, W. H., SAMANI, N. J. et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* **447** 661–678.

CARVALHO, L. and LAWRENCE, C. (2008). Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc. Natl. Acad. Sci. USA* **105** 3209–3214.

Key words and phrases. Large p small n, hierarchical Bayes, Pólya–Gamma latent variable.

COWLES, M. K. and CARLIN, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *J. Amer. Statist. Assoc.* **91** 883–904. MR1395755

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39** 1–38. MR0501537

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

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). A note on the group lasso and a sparse group lasso. Technical report, Stanford Univ., Stanford, CA. Available at arXiv:1001.0736.

GELFAND, A. E. and GHOSH, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85** 1–11. MR1627258

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GRABIEC, A. M., ANGIOLILLI, C., HARTKAMP, L. M., VAN BAARSEN, L. G., TAK, P. P. and REEDQUIST, K. A. (2014). JNK-dependent downregulation of FoxO1 is required to promote the survival of fibroblast-like synoviocytes in rheumatoid arthritis. *Annals of the Rheumatic Diseases* **74** annrheumdis–2013.

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

HABIER, D., FERNANDO, R., KIZILKAYA, K. and GARRIC, D. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* **12** 186.

HAMADA, M. and ASAI, K. (2012). A classification of bioinformatics algorithms from the viewpoint of maximizing expected accuracy (MEA). *J. Comput. Biol.* **19** 532–549. MR2925546

HAUPT, J., CASTRO, R. M. and NOWAK, R. (2011). Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Trans. Inform. Theory* **57** 6222–6235. MR2857969

HEARD, E., TISHKOFF, S., TODD, J. A., VIDAL, M., WAGNER, G. P., WANG, J., WEIGEL, D. and YOUNG, R. (2010). Ten years of genetics and genomics: What have we achieved and where are we heading? *Nat. Rev. Genet.* **11** 723–733.

HOERL, A. and KENNARD, R. (1970). Ridge regression—Applications to nonorthogonal problems. *Technometrics* **12** 69–82.

HOFFMAN, G. E., LOGSDON, B. A. and MEZEY, J. G. (2013). Puma: A unified framework for penalized multiple regression analysis of gwas data. *PLoS Comput. Biol.* **9** e1003101.

IOANNIDIS, J. P., THOMAS, G. and DALY, M. J. (2009). Validating, augmenting and refining genome-wide association signals. *Nat. Rev. Genet.* **10** 318–329.

ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158

JOHNSTON, I., HANCOCK, T., MAMITSUKA, H. and CARVALHO, L. (2016). Supplement to "Gene-proximity models for genome-wide association studies." DOI:10.1214/16-AOAS907SUPP.

JORGENSON, E. and WITTE, J. S. (2006). A gene-centric approach to genome-wide association studies. *Nat. Rev. Genet.* **7** 885–891.

KOOPERBERG, C., LEBLANC, M. and OBENCHAIN, V. (2010). Risk prediction using genome-wide association studies. *Genetic Epidemiology* **34** 643–652.

MACCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models* **37**. Chapman and Hall/CRC press, London.

MALACARDS (2014). Genes related to rheumatoid arthritis. Available at http://www.malacards.org/card/rheumatoid_arthritis. [Online. accessed 2014-10-01].

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models*. Chapman & Hall, London. MR0727836

MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503

MICHOU, L., LASBLEIZ, S., RAT, A.-C., MIGLIORINI, P., BALSA, A., WESTHOVENS, R., BARRERA, P., ALVES, H., PIERLOT, C., GLIKMANS, E. et al. (2007). Linkage proof for ptpn22, a rheumatoid arthritis susceptibility gene and a human autoimmunity gene. *Proc. Natl. Acad. Sci. USA* **104** 1649–1654.

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. MR0997578

PENG, B., ZHU, D., ANDER, B. P., ZHANG, X., XUE, F., SHARP, F. R. and YANG, X. (2013). An integrative framework for Bayesian variable selection with informative priors for identifying genes and pathways. *PloS One* **8** e67672.

PETERSEN, K. B. and PEDERSEN, M. S. (2012). The matrix cookbook. Technical University of Denmark.

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712

PRITCHARD, J. and PRZEWORSKI, M. (2001). Linkage disequilibrium in humans: Models and data. *American Journal of Human Genetics* **69** 1–14.

STEPHENS, M. and BALDING, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10** 681–690.

TECHNOLOGY DEPARTMENT CARNEGIE LIBRARY OF PITTSBURGH (2002). In The Handy Science Answer Book. Visible Ink Press.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641

WANG, W. Y., BARRATT, B. J., CLAYTON, D. G. and TODD, J. A. (2005). Genome-wide association studies: Theoretical and practical concerns. *Nat. Rev. Genet.* **6** 109–118.

WEST, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) 733–742. Oxford Univ. Press, New York. MR2003537

WHITTEMORE, A. S. (2007). A Bayesian false discovery rate for multiple testing. *J. Appl. Stat.* **34** 1–9. MR2345755

WIGGINTON, J. E., CUTLER, D. J. and ABECASIS, G. R. (2005). A note on exact tests of Hardy-weinberg equilibrium. *The American Journal of Human Genetics* **76** 887–893.

WU, M. C., KRAFT, P., EPSTEIN, M. P., TAYLOR, D. M., CHANOCK, S. J., HUNTER, D. J. and LIN, X. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics* **86** 929–942.

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. *The American Journal of Human Genetics* **89** 82–93.

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

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1245–1285
DOI: 10.1214/16-AOAS910

COMPARED TO WHAT? VARIATION IN THE IMPACTS OF EARLY CHILDHOOD EDUCATION BY ALTERNATIVE CARE TYPE[1]

BY AVI FELLER*, TODD GRINDAL†, LUKE MIRATRIX‡ AND LINDSAY C. PAGE§

University of California, Berkeley, Abt Associates†, Harvard University‡ and University of Pittsburgh§*

Early childhood education research often compares a group of children who receive the intervention of interest to a group of children who receive care in a range of different care settings. In this paper, we estimate differential impacts of an early childhood intervention by alternative care type, using data from the Head Start Impact Study, a large-scale randomized evaluation. To do so, we utilize a Bayesian principal stratification framework to estimate separate impacts for two types of Compliers: those children who would otherwise be in other center-based care when assigned to control and those who would otherwise be in home-based care. We find strong, positive short-term effects of Head Start on receptive vocabulary for those Compliers who would otherwise be in home-based care. By contrast, we find no meaningful impact of Head Start on vocabulary for those Compliers who would otherwise be in other center-based care. Our findings suggest that alternative care type is a potentially important source of variation in early childhood education interventions.

REFERENCES

ABADIE, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *J. Econometrics* **113** 231–263. MR1960380

ADMINISTRATION FOR CHILDREN AND FAMILIES (2014). Head Start Program Facts, Fiscal Year 2013. Available at https://eclkc.ohs.acf.hhs.gov/hslc/data/factsheets/docs/hs-program-fact-sheet-2013.pdf.

ANGRIST, J. D. (2004). Treatment effect heterogeneity in theory and practice. *Econ. J.* **114** C52–C83.

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

ANGRIST, J. D. and PISCHKE, J.-S. (2008). *Mostly Harmless Econometrics*: *An Empiricist's Companion*. Princeton University Press, Princeton, NJ.

BAFUMI, J. and GELMAN, A. E. (2006). Fitting multilevel models when predictors and group effects correlate. Unpublished manuscript.

BARNARD, J., FRANGAKIS, C. E., HILL, J. L. and RUBIN, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *J. Amer. Statist. Assoc.* **98** 299–323. MR1995712

Key words and phrases. Principal stratification, early childhood education, treatment effect variation, Head Start.

BARNETT, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *Future Child.* **5** 25.

BARNETT, W. S. (2011). Effectiveness of early educational intervention. *Science* **333** 975–978.

BARNETT, W. S. and HASKINS, R. (2010). *Investing in Young Children*: *New Directions in Federal Preschool and Early Childhood Policy*. The Brookings Institute, Washington, DC.

BARNETT, W. S., CAROLAN, M. E., SQUIRES, J. H. and BROWN, K. C. (2014). State of Preschool 2013: First Look. U.S. Dept. Education, National Center for Education Statistics, Washington, DC.

BASSOK, D., FITZPATRICK, M. and LOEB, S. (2013). Does state preschool crowd-out private provision? The impact of universal preschool on the childcare sector in Oklahoma and Georgia. NBER Working Paper 18605.

BITLER, M., GELBACH, J. and HOYNES, H. (2003). What mean impacts miss: Distributional effects of welfare reform experiments. *Am. Econ. Rev.* **96** 988–1012.

BITLER, M., HOYNES, H. and DOMINA, T. (2014). Experimental evidence on distributional effects of Head Start. Working paper.

BLOOM, H. S. and UNTERMAN, R. (2014). Can small high schools of choice improve educational prospects for disadvantaged students? *J. Policy Anal. Manage.* **33** 290–319.

BLOOM, H. S. and WEILAND, C. (2014). To what extent do the effects of Head Start on enrolled children vary across sites? Working paper.

BORDES, L., MOTTELET, S. and VANDEKERKHOVE, P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34** 1204–1232. MR2278356

BURGESS, K., CHIEN, N., MORRISSEY, T. and SWENSON, K. (2014). Trends in the use of early care and education, 1995–2011: Descriptive analysis of child care arrangements from national survey data. Report from the Office of the Assistant Secretary for Plannng and Evaluation, US Department of Health and Human Services.

CARNEIRO, P. and GINJA, R. (2014). Long term impacts of compensatory preschool on health and behavior: Evidence from Head Start. *Am. Econ. J. Appl. Econ.* **6** 135–173.

CASCIO, E. U. and SCHANZENBACH, D. W. (2013). The impacts of expanding access to high-quality preschool education. In *Brookings Papers on Economic Activity* 127–192. Brookings Institution, Washington, DC.

WESTINGHOUSE LEARNING CORPORATION (1969). *The Impact of Head Start*: *An Evaluation of the Effects of Head Start on Children's Cognitive and Affective Development*, *Vol.* 1: *Report to the Office of Economic Opportunity*. Westinghouse Learning Corporation and Ohio Univ., Athens, Ohio.

COULSON, A. J. (2013). Preschool's Anvil Chorus. Cato Institute, Washington, DC.

COX, D. R. and DONNELLY, C. A. (2011). *Principles of Applied Statistics*. Cambridge Univ. Press, Cambridge. MR2817147

CURRIE, J. and THOMAS, D. (1995). Does Head Start make a difference? *Am. Econ. Rev.* **85** 341–364.

DAY, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56** 463–474. MR0254956

DEMING, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *Am. Econ. J. Appl. Econ.* **1** 111–134.

DING, P., GENG, Z., YAN, W. and ZHOU, X.-H. (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *J. Amer. Statist. Assoc.* **106** 1578–1591. MR2896858

DUNCAN, G. J. and MAGNUSON, K. (2013). Investing in preschool programs. *J. Econ. Perspect.* **27** 109–132.

ELANGO, S., GARCÍA, J. L., HECKMAN, J. J. and HOJMAN, A. (2015). Early childhood education. Technical report, National Bureau of Economic Research Working Paper No. 21766.

FELLER, A. (2015). Essays in public policy and causal inference. Ph.D. thesis, Harvard Univ., Cambridge, MA.

FELLER, A., GREIF, E., MIRATRIX, L. and PILLAI, N. (2016). Principal stratification in the Twilight Zone: Weakly separated components in finite mixture models. Available at arXiv:1602.06595.

FELLER, A., GRINDAL, T., MIRATRIX, L. and PAGE, L. C. (2016). Supplement to "Compared to what? Variation in the impacts of early childhood education by alternative care type." DOI:10.1214/16-AOAS910SUPP.

FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York. MR2265601

FRUMENTO, P., MEALLI, F., PACINI, B. and RUBIN, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *J. Amer. Statist. Assoc.* **107** 450–466. MR2980057

FRUMENTO, P., MEALLI, F., PACINI, B. and RUBIN, D. B. (2016). The fragility of standard inferential approaches in principal stratification models relative to direct likelihood approaches. *Stat. Anal. Data Min.* **9** 58–70. MR3465093

FRYER, R. G. and LEVITT, S. D. (2004). Understanding the black–white test score gap in the first two years of school. *Rev. Econ. Stat.* **86** 447–464.

GALLOP, R., SMALL, D. S., LIN, J. Y., ELLIOTT, M. R., JOFFE, M. and TEN HAVE, T. R. (2009). Mediation analysis with principal stratification. *Stat. Med.* **28** 1108–1130. MR2662200

GARCES, E., THOMAS, D. and CURRIE, J. (2002). Longer-term effects of Head Start. *Am. Econ. Rev.* **92** 999–1012.

GELBER, A. and ISEN, A. (2013). Children's schooling and parents' behavior: Evidence from the Head Start Impact Study. *J. Public Econ.* **101** 25–38.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*. CRC press, Boca Raton, FL.

GIBBS, C., LUDWIG, J. and MILLER, D. L. (2013). Does Head Start do any lasting good? In *The War on Poverty: A 50-Year Retrospective* (M. J. Bailey and Sheldon Danziger, eds.). Russell Sage Foundation, New York.

GORMLEY, W. T. (2007). Early childhood care and education: Lessons and puzzles. *J. Policy Anal. Manage.* **26** 633–671.

GORMLEY, W. T., PHILLIPS, D., ADELSTEIN, S. and SHAW, C. (2010). Head Start's comparative advantage: Myth or reality? *Policy Stud. J.* **38** 397–418.

GRIFFIN, B. A., MCCAFFREY, D. F. and MORRAL, A. R. (2008). An application of principal stratification to control for institutionalization at follow-up in studies of substance abuse treatment programs. *Ann. Appl. Stat.* **2** 1034–1055. MR2516803

HALL, P. and ZHOU, X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.* **31** 201–224. MR1962504

HECKMAN, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science* **312** 1900–1902.

HECKMAN, J., HOHMANN, N., SMITH, J. and KHOO, M. (2000). Substitution and dropout bias in social experiments: A study of an influential social experiment. *Q. J. Econ.* **115** 651–694.

HILL, J., WALDFOGEL, J. and BROOKS-GUNN, J. (2002). Differential effects of high-quality child care. *J. Policy Anal. Manage.* **21** 601–627.

HIRANO, K., IMBENS, G. W., RUBIN, D. B. and ZHOU, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1** 69–88.

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

HUNTER, D. R., WANG, S. and HETTMANSPERGER, T. P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35** 224–251. MR2332275

HUSTON, A. C., CHANG, Y. E. and GENNETIAN, L. (2002). Family and individual predictors of child care use by low-income families in different policy contexts. *Early Child. Res. Q.* **17** 441–469.

IMAI, K., KING, G. and STUART, E. A. (2008). Misunderstanding between experimentalists and observationalists about causal inference. *J. Roy. Statist. Soc. Ser. A* **171** 481–502. MR2427345

IMBENS, G. W. and RUBIN, D. B. (1997a). Estimating outcome distributions for compliers in instrumental variables models. *Rev. Econ. Stud.* **64** 555–574. MR1485828

IMBENS, G. W. and RUBIN, D. B. (1997b). Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.* **25** 305–327. MR1429927

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951

JENKINS, J. M., FARKAS, G., DUNCAN, G. J., BURCHINAL, M. and VANDELL, D. L. (2014). Head Start at ages 3 and 4 versus Head Start followed by state pre-k: Which is more effective? Working paper.

JIN, H. and RUBIN, D. B. (2009). Public schools versus private schools: Causal inference with partial compliance. *J. Educ. Behav. Stat.* **34** 24–45.

JO, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics* **27** 385–409.

JO, B. and MUTHÉN, B. O. (2001). Modeling of intervention effects with noncompliance: A latent variable approach for randomized trials. In *New Developments and Techniques in Structural Equation Modeling* (G. A. Marcoulides and R. E. Schumacker, eds.) 57–87. Erlbaum Associates, Mahwah, NJ.

JO, B. and STUART, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Stat. Med.* **28** 2857–2875. MR2750169

JOFFE, M. M., SMALL, D. and HSU, C.-Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statist. Sci.* **22** 74–97. MR2408662

KAGAN, S. L. (1991). Examining profit and nonprofit child care: An odyssey of quality and auspices. *J. Soc. Issues* **47** 87–104.

KLINE, P. and WALTERS, C. (2016). Evaluating public programs with close substitutes: The case of Head Start. *Q. J. Econ.* To appear. DOI:10.1093/qje/qjw027.

KLING, J. R., LIEBMAN, J. B. and KATZ, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica* **75** 83–119.

LEAK, J., DUNCAN, G. J., LI, W., MAGNUSON, K. A., SCHINDLER, H. and YOSHIKAWA, H. (2010). Is timing everything? How early childhood education program impacts vary by starting age, program duration and time since the end of the program. Working paper.

LUDWIG, J. and MILLER, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Q. J. Econ.* **122** 159–208.

LUDWIG, J. and PHILLIPS, D. A. (2010). Leave no (young) child behind: Prioritizing access in early childhood education. In *Investing in Young Children: New Directions in Federal Preschool and Early Childhood Policy* (R. Haskin and W. S. Barnett, eds.). Brookings and NIEER.

MAGNUSON, K. A., RUHM, C. and WALDFOGEL, J. (2007). The persistence of preschool effects: Do subsequent classroom experiences matter?. *Early Child. Res. Q.* **22** 18–38.

MATTEI, A., LI, F. and MEALLI, F. (2013). Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program. *Ann. Appl. Stat.* **7** 2336–2360. MR3161725

MCCOY, D. C., CONNORS, M. C., MORRIS, P. A., YOSHIKAWA, H. and FRIEDMAN-KRAUSS, A. H. (2015). Neighborhood economic disadvantage and children's cognitive and social-emotional development: Exploring Head Start classroom quality as a mediating mechanism. *Early Childhood Research Quarterly* **32** 150–159.

MEALLI, F. and PACINI, B. (2008). Comparing principal stratification and selection models in parametric causal inference with nonignorable missingness. *Comput. Statist. Data Anal.* **53** 507–516. MR2649105

MEALLI, F. and PACINI, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *J. Amer. Statist. Assoc.* **108** 1120–1131. MR3174688

MILLER, E. B., FARKAS, G., VANDELL, D. L. and DUNCAN, G. J. (2014). Do the effects of Head Start vary by parental preacademic stimulation? *Child Dev.* **85** 1385–1400.

MORRIS, J. R. and HELBURN, S. W. (2000). Child care center quality differences: The role of profit status, client preferences, and trust. *Nonprofit Volunt. Sect. Q.* **29** 377–399.

NATIONAL FORUM ON EARLY CHILDHOOD POLICY AND PROGRAMS (2010). Understanding the Head Start impact study. Available at http://developingchild.harvard.edu.

PAGE, L. C. (2012). Principal stratification as a framework for investigating mediational processes in experimental settings. *Journal of Research on Educational Effectiveness* **5** 215–244.

PAGE, L. C., FELLER, A., GRINDAL, T., MIRATRIX, L. and SOMERS, M. A. (2015). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *Am J. Eval.* **36** 514–531.

PEARSON, K. (1894). Contributions to the mathematical theory of evolution. *Philos. Trans. R. Soc. Lond., A* **185** 71–110.

PUMA, M., BELL, S. H., COOK, R., HEID, C. and SHAPIRO, G. (2010a). Head Start impact study. Final report, HHS, Administration for Children and Families.

PUMA, M., BELL, S. H., COOK, R., HEID, C. and SHAPIRO, G. (2010b). Head Start impact study. Technical report, HHS, Administration for Children and Families.

RAUDENBUSH, S. W. (2015). Estimation of means and covariance components in multi-site randomized trials. Unpublished manuscript.

RAUDENBUSH, S. W., REARDON, S. F. and NOMI, T. (2012). Statistical analysis for multisite trials using instrumental variables with random coefficients. *Journal of Research on Educational Effectiveness* **5** 303–332.

REARDON, S. F. and RAUDENBUSH, S. W. (2013). Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociol. Methods Res.* **42** 143–163. MR3190727

ROMANO, E., BABCHISHIN, L., PAGANI, L. S. and KOHEN, D. (2010). School readiness and later achievement: Replication and extension using a nationwide Canadian survey. *Dev. Psychol.* **46** 995–1007.

ROSE, K. K. and ELICKER, J. (2010). Maternal child care preferences for infants, toddlers, and preschoolers: The disconnect between policy and preference in the USA. *Community Work Fam.* **13** 205–229.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

RUBIN, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test". *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

SCHMIT, S., MATTHEWS, H., SMITH, S. and ROBBINS, T. (2013). Investing in young children: A fact sheet on early care and education participation, access, and quality. Fact sheet, New York, NY: National Center for Children in Poverty, Washington, DC: Center for Law and Social Policy.

SCHOCHET, P. Z. (2013). Student mobility, dosage, and principal stratification in school-based RCTs. *J. Educ. Behav. Stat.* **38** 323–354.

SCHOCHET, P. Z. and BURGHARDT, J. (2007). Using propensity scoring to estimate program-related subgroup impacts in experimental program evaluations. *Eval. Rev.* **31** 95–120.

SCHOCHET, P. Z., BURGHARDT, J. and MCCONNELL, S. (2008). Does job corps work? Impact findings from the national job corps study. *Am. Econ. Rev.* **98** 1864–1886.

SCHOCHET, P., PUMA, M. and DEKE, J. (2014). Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods (NCEE 2014–4017), Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development.

SCOTT-CLAYTON, J. and MINAYA, V. (2014). Should student employment be subsidized? Conditional counterfactuals and the outcomes of work-study participation. Working Paper w20329, National Bureau of Economic Research.

SHAGER, H. M., SCHINDLER, H. S., MAGNUSON, K. A., DUNCAN, G. J., YOSHIKAWA, H. and HART, C. M. D. (2013). Can research design explain variation in Head Start research results? A meta-analysis of cognitive and achievement outcomes. *Educ. Eval. Policy Anal.* **35** 76–95.

SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. MR1092986

STAN DEVELOPMENT TEAM (2014). Stan: A C++ library for probability and sampling, Version 2.3.

WALTERS, C. R. (2015). Inputs in the production of early childhood human capital: Evidence from Head Start. *Am. Econ. J. Appl. Econ.* **7** 76–102.

WHITEHURST, G. J. (2013a). Obama's preschool plan. Brookings Institution.

WHITEHURST, G. J. (2013b). Can we be hard-headed about preschool? A look at Head Start. Brookings Institution.

ZHAI, F., BROOKS-GUNN, J. and WALDFOGEL, J. (2011). Head Start and urban children's school readiness: A birth cohort study in 18 cities. *Dev. Psychol.* **47** 134–152.

ZHAI, F., BROOKS-GUNN, J. and WALDFOGEL, J. (2014). Head Start's impact is contingent on alternative type of care in comparison group. *Dev. Psychol.* **50** 2572–2586.

ZHANG, J. L. and RUBIN, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *J. Educ. Behav. Stat.* **28** 353–368.

ZHANG, J. L., RUBIN, D. B. and MEALLI, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *J. Amer. Statist. Assoc.* **104** 166–176. MR2663040

ZIGLER, E. and MUENCHOW, S. (1992). *Head Start*: *The Inside Story of America's Most Successful Educational Experiment*. Basic Books.

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1286–1316
DOI: 10.1214/16-AOAS931

NONSEPARABLE DYNAMIC NEAREST NEIGHBOR GAUSSIAN PROCESS MODELS FOR LARGE SPATIO-TEMPORAL DATA WITH AN APPLICATION TO PARTICULATE MATTER ANALYSIS

BY ABHIRUP DATTA*, SUDIPTO BANERJEE[†,1], ANDREW O. FINLEY[‡,2], NICHOLAS A. S. HAMM[§] AND MARTIJN SCHAAP[¶]

Johns Hopkins University, University of California, Los Angeles*†, *Michigan State University*‡, University of Twente*§ and TNO*¶

Particulate matter (PM) is a class of malicious environmental pollutants known to be detrimental to human health. Regulatory efforts aimed at curbing PM levels in different countries often require high resolution space–time maps that can identify red-flag regions exceeding statutory concentration limits. Continuous spatio-temporal Gaussian Process (GP) models can deliver maps depicting predicted PM levels and quantify predictive uncertainty. However, GP-based approaches are usually thwarted by computational challenges posed by large datasets. We construct a novel class of scalable Dynamic Nearest Neighbor Gaussian Process (DNNGP) models that can provide a sparse approximation to any spatio-temporal GP (e.g., with nonseparable covariance structures). The DNNGP we develop here can be used as a sparsity-inducing prior for spatio-temporal random effects in any Bayesian hierarchical model to deliver full posterior inference. Storage and memory requirements for a DNNGP model are linear in the size of the dataset, thereby delivering massive scalability without sacrificing inferential richness. Extensive numerical studies reveal that the DNNGP provides substantially superior approximations to the underlying process than low-rank approximations. Finally, we use the DNNGP to analyze a massive air quality dataset to substantially improve predictions of PM levels across Europe in conjunction with the LOTOS-EUROS chemistry transport models (CTMs).

REFERENCES

ALLCROFT, D. J. and GLASBEY, C. A. (2003). A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. *J. Roy. Statist. Soc. Ser. C* **52** 487–498. MR2012972

BAI, Y., SONG, P. X. K. and RAGHUNATHAN, T. E. (2012). Bayesian dynamic modeling for large space–time datasets using Gaussian predictive processes. *J. Roy. Statist. Soc. Ser. B* **74** 799–824.

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2014). Hierarchical Modeling and Analysis for Spatial Data, 2nd ed. Chapman & Hall, Boca Raton, FL.

BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 825–848. MR2523906

BEVILACQUA, M., GAETAN, C., MATEU, J. and PORCU, E. (2012). Estimating space and space–time covariance functions for large data sets: A weighted composite likelihood approach. *J. Amer. Statist. Assoc.* **107** 268–280. MR2949358

Key words and phrases. Nonseparable spatio-temporal models, scalable Gaussian process, nearest neighbors, Bayesian inference, Markov chain Monte Carlo, environmental pollutants.

BEVILACQUA, M., FASS, Ò. A., GAETAN, C., PORCU, E.and VELANDIA, D. (2015). Covariance tapering for multivariate Gaussian random fields estimation. *Stat. Methods Appl.* **25** 21–37.

BIRMILI, W., SCHEPANSKI, K., ANSMANN, A., SPINDLER, G., TEGEN, I., WEHNER, B., NOWAK, A., REIMER, E., MATTIS, I., MULLER, K., BRUGGEMANN, E., GNAUK, T., HERRMANN, H., WIEDENSOHLER, A., ALTHAUSEN, D., SCHLADITZ, A., TUCH, T. and LOSCHAU, G. (2008). A case of extreme particulate matter concentrations over central Europe caused by dust emitted over the southern Ukraine. *Atmos. Chem. Phys.* **8** 997–1016.

BRAUER, M., AMANN, M., BURNETT, R. T., COHEN, A., DENTENER, F., EZZATI, M., HENDERSON, S. B., KRZYZANOWSKI, M., MARTIN, R. V., VAN DINGENEN, R., VAN DONKELAAR, A. and THURSTON, G. D. (2011). Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution. *Environ. Sci. Technol.* **46** 652–660.

BRUNEKREEF, B. and HOLGATE, S. T. (2002). Air pollution and health. *Lancet* **360** 1233–1242.

CANDIANI, G., CARNEVALE, C., FINZI, G., PISONI, E. and VOLTA, M. (2013). A comparison of reanalysis techniques: Applying optimal interpolation and ensemble Kalman filtering to improve air quality monitoring at mesoscale. *Sci. Total Environ.* **458–460** 7–14.

CRAINICEANU, C. M., DIGGLE, P. J. and ROWLINGSON, B. (2008). Bivariate binomial spatial modeling of *Loa loa* prevalence in tropical Africa. *J. Amer. Statist. Assoc.* **103** 21–37. MR2420211

CRESSIE, N. and HUANG, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Amer. Statist. Assoc.* **94** 1330–1340. MR1731494

CRESSIE, N. and JOHANNESSON, G. (2008). Fixed rank kriging for very large spatial data sets. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 209–226. MR2412639

CRESSIE, N., SHI, T. and KANG, E. L. (2010). Fixed rank filtering for spatio-temporal data. *J. Comput. Graph. Statist.* **19** 724–745. MR2732500

CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ. MR2848400

DAGUM, L. and MENON, R. (1998). OpenMP: An industry standard API for shared-memory programming. *IEEE Comput. Sci. Eng.* **5** 46–55.

DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016a). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812.

DATTA, A., BANERJEE, S., FINLEY, A. O., HAMM, N. S. and SCHAAP, M. (2016b). Supplement to "Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis." DOI:10.1214/16-AOAS931SUPP.

DENBY, B., SCHAAP, M., SEGERS, A., BUILTJES, P. and HORALEK, J. (2008). Comparison of two data assimilation methods for assessing PM10 exceedances on the European scale. *Atmos. Environ.* **42** 7122–7134.

DENBY, B., SUNDVOR, I., CASSIANI, M., DE SMET, P., DE LEEUW, F. and HORALEK, J. (2010). Spatial mapping of ozone and SO2 trends in Europe. *Sci. Total Environ.* **408** 4795–4806.

DU, J., ZHANG, H. and MANDREKAR, V. S. (2009). Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *Ann. Statist.* **37** 3330–3361. MR2549562

EEFTENS, M., TSAI, M. Y., AMPE, C., ANWANDER, B., BEELEN, R., BELLANDER, T., CESARONI, G., CIRACH, M., CYRYS, J., DE HOOGH, K., DE NAZELLE, A., DE VOCHT, F., DECLERCQ, C., DEDELE, A., ERIKSEN, K., GALASSI, C., GRAZULEVICIENE, R., GRIVAS, G., HEINRICH, J., HOFFMANN, B., IAKOVIDES, M., INEICHEN, A., KATSOUYANNI, K., KOREK, M., KRAMER, U., KUHLBUSCH, T., LANKI, T., MADSEN, C., MELIEFSTE, K., MOLTER, A., MOSLER, G., NIEUWENHUIJSEN, M., OLDENWENING, M., PENNANEN, A., PROBST-HENSCH, N., QUASS, U., RAASCHOU-NIELSEN, O., RANZI, A., STEPHANOU, E., SUGIRI, D., UDVARDY, O., VASKOEVI, E., WEINMAYR, G., BRUNEKREEF, B. and HOEK, G. (2012). Spatial variation of PM2.5, PM10, PM2.5 absorbance and PMcoarse concentrations between and within 20 European study areas and the relationship with NO2—results of the ESCAPE project. *Atmos. Environ.* **62** 303–317.

EIDSVIK, J., SHABY, B. A., REICH, B. J., WHEELER, M. and NIEMI, J. (2014). Estimation and prediction in spatial models with block composite likelihoods. *J. Comput. Graph. Statist.* **23** 295–315. MR3215812

EUROPEAN COMMISSION (2015). European Union Air Quality Standards. Available at http://ec.europa.eu/environment/air/quality/standards.htm.

FINLEY, A. O., BANERJEE, S. and GELFAND, A. E. (2012). Bayesian dynamic modeling for large space–time datasets using Gaussian predictive processes. *J. Geogr. Syst.* **14** 29–47.

FINLEY, A. O., BANERJEE, S. and MCROBERTS, R. E. (2009). Hierarchical spatial models for predicting tree species assemblages across large domains. *Ann. Appl. Stat.* **3** 1052–1079. MR2750386

FLEMMING, J., INNESS, A., FLENTJE, H., HUIJNEN, V., MOINAT, P., SCHULTZ, M. G. and STEIN, O. (2009). Coupling global chemistry transport models to ECMWF's integrated forecast system. *Geosci. Model Dev.* **2** 253–265.

FURRER, R., GENTON, M. G. and NYCHKA, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15** 502–523. MR2291261

GELFAND, A. E., BANERJEE, S. and GAMERMAN, D. (2005). Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics* **16** 465–479. MR2147537

GELFAND, A. E. and GHOSH, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85** 1–11. MR1627258

GELFAND, A. E., DIGGLE, P. J., FUENTES, M. and GUTTORP, P., eds. (2010). *Handbook of Spatial Statistics. Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR2761512

GNEITING, T. (2002). Nonseparable, stationary covariance functions for space–time data. *J. Amer. Statist. Assoc.* **97** 590–600. MR1941475

GNEITING, T., GENTON, M. G. and GUTTORP, P. (2007). Geostatistical space–time models, stationarity, separability and full symmetry. In *Statistics of SpatioTemporal Systems* (B. Finkenstaedt, L. Held and V. Isham, eds.) 151–175. Chapman & Hall, London.

GNEITING, T. and GUTTORP, P. (2010). Continuous parameter spatio-temporal processes. In *Handbook of Spatial Statistics. Handb. Mod. Stat. Methods* (A. E. Gelfand, P. Diggle, M. Fuentes and P. Guttorp, eds.) 427–436. CRC Press, Boca Raton, FL. MR2730958

GRÄLER, B., GERHARZ, L. and PEBESMA, E. (2011). Spatio-temporal analysis and interpolation of PM10 measurements in Europe. ETC/ACM Technical Paper 2011/10, European Topic Centre on Air Pollution and Climate Change Mitigation, Bilthoven, The Netherlands.

GRAMACY, R. B. and APLEY, D. W. (2015). Local Gaussian process approximation for large computer experiments. *J. Comput. Graph. Statist.* **24** 561–578. MR3357395

HAMM, N. A. S., FINLEY, A. O., SCHAAP, M. and STEIN, A. (2015). A spatially varying coefficient model for mapping PM10 air quality at the European scale. *Atmos. Environ.* **102** 393–405.

HENDRIKS, C., KRANENBURG, R., KUENEN, J., VAN GIJLSWIJK, R., KRUIT, R. W., SEGERS, A., VAN DER GON, H. D. and SCHAAP, M. (2013). The origin of ambient particulate matter concentrations in the Netherlands. *Atmospheric Environment* **69** 289–303.

HIGDON, D. (2001). Space and space time modeling using process convolutions. Technical report, Institute of Statistics and Decision Sciences, Duke Univ., Durham, NC.

HOEK, G., KRISHNAN, R. M., BEELEN, R., PETERS, A., OSTRO, B., BRUNEKREEF, B. and KAUFMAN, J. D. (2013). Long-term air pollution exposure and cardio-respiratory mortality: A review. *Environ. Health* **12** 43.

INTEL (2015). Math Kernel Library. Available at http://developer.intel.com/software/products/mkl/.

JONES, R. H. and ZHANG, Y. (1997). Models for continuous stationary space–time processes. In *Modelling Longitudinal and Spatially Correlated Data* (T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren and R. D. Wolfinger, eds.) 289–298. Springer, New York.

KAMMANN, E. E. and WAND, M. P. (2003). Geoadditive models. *J. Roy. Statist. Soc. Ser. C* **52** 1–18. MR1963210

KATZFUSS, M. (2016). A multi-resolution approximation for massive spatial datasets. *J. Amer. Statist. Assoc*. Available at arXiv:1507.04789.

KATZFUSS, M. and CRESSIE, N. (2012). Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmetrics* **23** 94–107. MR2873787

KAUFMAN, C. G., SCHERVISH, M. J. and NYCHKA, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *J. Amer. Statist. Assoc*. **103** 1545–1555. MR2504203

KYRIAKIDIS, P. C. and JOURNEL, A. G. (1999). Geostatistical space–time models: A review. *Math. Geol*. **31** 651–684. MR1694654

LLOYD, C. D. and ATKINSON, P. M. (2004). Increased accuracy of geostatistical prediction of nitrogen dioxide in the United Kingdom with secondary data. *International Journal of Applied Earth Observation and Geoinformation* **5** 293–305.

LOOMIS, D., GROSSE, Y., LAUBY-SECRETAN, B., EL GHISSASSI, F., BOUVARD, V., BENBRAHIM-TALLAA, L., GUHA, N., BAAN, R., MATTOCK, H. and STRAIF, S. (2013). The carcinogenicity of outdoor air pollution. *Lancet Oncol*. **14** 1262–1263.

MANDERS, A. M. M., SCHAAP, M. and HOOGERBRUGGE, R. (2009). Testing the capability of the chemistry transport model LOTOS-EUROS to forecast PM10 levels in the Netherlands. *Atmos. Environ*. **43** 4050–4059.

MUES, A., KUENEN, J., HENDRIKS, C., MANDERS, A., SEGERS, A., SCHOLZ, Y., HUEGLIN, C., BUILTJES, P. and SCHAAP, M. (2014). Sensitivity of air pollution simulations with LOTOS-EUROS to the temporal distribution of anthropogenic emissions. *Atmos. Chem. Phys*. **14** 939–955.

OMIDI, M. and MOHAMMADZADEH, M. (2015). A new method to build spatio-temporal covariance functions: Analysis of ozone data. *Statist. Papers* 1–15.

PFEIFER, P. E. and DEUTSCH, S. J. (1980a). Independence and sphericity tests for the residuals of space–time ARMA models. *Comm. Statist. Simulation Comput*. **9** 533–549.

PFEIFER, P. E. and DEUTSCH, S. J. (1980b). Stationarity and invertibility regions for low order STARMA models. *Comm. Statist. Simulation Comput*. **9** 551–562.

POULIOT, G., PIERCE, T., VAN DER GON, H. D., SCHAAP, M., MORAN, M. and NOPMONGCOL, U. (2012). Comparing emission inventories and model-ready emission datasets between Europe and North America for the AQMEII project. *Atmos. Environ*. **53** 4–14.

R'HONI, Y., CLARISSE, L., CLERBAUX, C., HURTMANS, D., DUFLOT, V., TURQUETY, S., NGADI, Y. and COHEUR, P. F. (2013). Exceptional emissions of NH3 and HCOOH in the 2010 Russian wildfires. *Atmos. Chem. Phys*. **13** 4171–4181.

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2005). *Gaussian Processes for Machine Learning*, 1st ed. MIT Press, Cambridge, MA.

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields*: *Theory and Applications*. *Monographs on Statistics and Applied Probability* **104**. Chapman & Hall, Boca Raton, FL. MR2130347

SANG, H. and HUANG, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **74** 111–132. MR2885842

SCHAAP, M., TIMMERMANS, R. M. A., ROEMER, M., BOERSEN, G. A. C., BUILTJES, P., SAUTER, F., VELDERS, G. and BECK, J. (2008). The LOTOS-EUROS model: Description, validation and latest developments. *Int. J. Environ. Pollut*. **32** 270–290.

SHABY, B. and RUPPERT, D. (2012). Tapered covariance: Bayesian estimation and asymptotics. *J. Comput. Graph. Statist*. **21** 433–452. MR2945475

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol*. **64** 583–639. MR1979380

STEIN, M. L. (2005). Space–time covariance functions. *J. Amer. Statist. Assoc*. **100** 310–321. MR2156840

STEIN, M. L. (2007). Spatial variation of total column ozone on a global scale. *Ann. Appl. Stat.* **1** 191–210. MR2393847

STEIN, M. L. (2008). A modeling approach for large spatial datasets. *J. Korean Statist. Soc.* **37** 3–10. MR2420389

STEIN, M. L. (2013). On a class of space–time intrinsic random functions. *Bernoulli* **19** 387–408. MR3037158

STEIN, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spat. Stat.* **8** 1–19. MR3326818

STEIN, M. L., CHI, Z. and WELTY, L. J. (2004). Approximating likelihoods for large spatial data sets. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 275–296. MR2062376

STERN, R., BUILTJES, P., SCHAAP, M., TIMMERMANS, R., VAUTARD, R., HODZIC, A., MEMMESHEIMER, M., FELDMANN, H., RENNER, E., WOLKE, R. and KERSCHBAUMER, A. (2008). A model inter-comparison study focussing on episodes with elevated PM10 concentrations. *Atmos. Environ.* **42** 4567–4588.

STOFFER, D. S. (1986). Estimation and identification of space–time ARMAX models in the presence of missing data. *J. Amer. Statist. Assoc.* **81** 762–772. MR0860510

STROUD, J. R., MÜLLER, P. and SANSÓ, B. (2001). Dynamic models for spatiotemporal data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 673–689. MR1872059

VAN DE KASSTEELE, J. and STEIN, A. (2006). A model for external drift kriging with uncertain covariates applied to air quality measurements and dispersion model output. *Environmetrics* **17** 309–322. MR2239674

VECCHIA, A. V. (1988). Estimation and model identification for continuous spatial processes. *J. Roy. Statist. Soc. Ser. B* **50** 297–312. MR0964183

VECCHIA, A. V. (1992). A new method of prediction for spatial regression models with correlated errors. *J. Roy. Statist. Soc. Ser. B* **54** 813–830. MR1185224

XU, G., LIANG, F. and GENTON, M. G. (2015). A Bayesian spatio-temporal geostatistical model with an auxiliary lattice for large datasets. *Statist. Sinica.* **25** 61–79.

YENIAY, Ö. and GÖKTAŞ, A. (2002). A comparison of partial least squares regression with other prediction methods. *Hacet. J. Math. Stat.* **31** 99–111. MR1987061

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1317–1347
DOI: 10.1214/16-AOAS934

PARALLEL PARTIAL GAUSSIAN PROCESS EMULATION FOR COMPUTER MODELS WITH MASSIVE OUTPUT[1]

BY MENGYANG GU* AND JAMES O. BERGER*,†

*Duke University** *and King Abdulaziz University, Jeddah, Saudi Arabia*†

We consider the problem of emulating (approximating) computer models (simulators) that produce massive output. The specific simulator we study is a computer model of volcanic pyroclastic flow, a single run of which produces up to 10^9 outputs over a space–time grid of coordinates. An emulator (essentially a statistical model of the simulator—we use a Gaussian Process) that is computationally suitable for such massive output is developed and studied from practical and theoretical perspectives. On the practical side, the emulator does unexpectedly well in predicting what the simulator would produce, even better than much more flexible and computationally intensive alternatives. This allows the attainment of the scientific goal of this work, accurate assessment of the hazards from pyroclastic flows over wide spatial domains. Theoretical results are also developed that provide insight into the unexpected success of the massive emulator. Generalizations of the emulator are introduced that allow for a nugget, which is useful for the application to hazard assessment.

REFERENCES

ANDRIANAKIS, I. and CHALLENOR, P. G. (2012). The effect of the nugget on Gaussian process emulators of computer models. *Comput. Statist. Data Anal.* **56** 4215–4228. MR2957866

BASTOS, L. S. and O'HAGAN, A. (2009). Diagnostics for Gaussian process emulators. *Technometrics* **51** 425–438. MR2756478

BAYARRI, M. J., BERGER, J. O., PAULO, R., SACKS, J., CAFEO, J. A., CAVENDISH, J., LIN, C.-H. and TU, J. (2007a). A framework for validation of computer models. *Technometrics* **49** 138–154. MR2380530

BAYARRI, M. J., BERGER, J. O., CAFEO, J., GARCIA-DONATO, G., LIU, F., PALOMO, J., PARTHASARATHY, R. J., PAULO, R., SACKS, J. and WALSH, D. (2007b). Computer model validation with functional output. *Ann. Statist.* **35** 1874–1906. MR2363956

BAYARRI, M. J., BERGER, J. O., CALDER, E. S., DALBEY, K., LUNAGOMEZ, S., PATRA, A. K., PITMAN, E. B., SPILLER, E. T. and WOLPERT, R. L. (2009). Using statistical and computer models to quantify volcanic hazards. *Technometrics* **51** 402–413. MR2756476

BAYARRI, M. J., BERGER, J. O., CALDER, E. S., PATRA, A. K., PITMAN, E. B., SPILLER, E. T. and WOLPERT, R. L. (2015). Probabilistic quantification of hazards: A methodology using small ensembles of physics-based simulations and statistical surrogates. *Int. J. Uncertain. Quantif.* **5** 297–325. MR3413743

BERGER, J. O., DE OLIVEIRA, V. and SANSÓ, B. (2001). Objective Bayesian analysis of spatially correlated data. *J. Amer. Statist. Assoc.* **96** 1361–1374. MR1946582

Key words and phrases. Gaussian process, computer model emulation, space–time coordinate, objective Bayesian analysis.

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

CONTI, S. and O'HAGAN, A. (2010). Bayesian emulation of complex multi-output and dynamic computer models. *J. Statist. Plann. Inference* **140** 640–651. MR2558393

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276. MR0400509

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York. MR1239641

FORRESTER, A., SOBESTER, A. and KEANE, A. (2008). *Engineering Design Via Surrogate Modelling*: *A Practical Guide*. Wiley, New York.

FRICKER, T. E., OAKLEY, J. E. and URBAN, N. M. (2013). Multivariate Gaussian process emulators with nonseparable covariance structures. *Technometrics* **55** 47–56. MR3038484

GELFAND, A. E., DIGGLE, P. J., FUENTES, M. and GUTTORP, P., eds. (2010). *Handbook of Spatial Statistics*. CRC Press, Boca Raton, FL. MR2761512

GU, M. (2016). Robust uncertainty quantification and scalable computation for computer models with massive output. Ph.D. thesis, Duke Univ.

GU, M. and BERGER, J. O. (2016). Supplement to "Parallel partial Gaussian process emulation for computer models with massive output." DOI:10.1214/16-AOAS934SUPP.

GUPTA, A. K. and NAGAR, D. K. (1999). *Matrix Variate Distributions*. CRC Press, Boca Raton.

HIGDON, D., GATTIKER, J., WILLIAMS, B. and RIGHTLEY, M. (2008). Computer model calibration using high-dimensional output. *J. Amer. Statist. Assoc.* **103** 570–583. MR2523994

IOOSS, B. and LEMAÎTRE, P. (2014). A review on global sensitivity analysis methods. Preprint. Available at arXiv:1404.2405.

KAUFMAN, C. G., SCHERVISH, M. J. and NYCHKA, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *J. Amer. Statist. Assoc.* **103** 1545–1555. MR2504203

KAUFMAN, C. G., BINGHAM, D., HABIB, S., HEITMANN, K. and FRIEMAN, J. A. (2011). Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology. *Ann. Appl. Stat.* **5** 2470–2492. MR2907123

KAZIANKA, H. and PILZ, J. (2012). Objective Bayesian analysis of spatial data with uncertain nugget and range parameters. *Canad. J. Statist.* **40** 304–327. MR2927748

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

KENNEDY, M., ANDERSON, C., O'HAGAN, A., LOMAS, M., WOODWARD, I., GOSLING, J. P. and HEINEMEYER, A. (2008). Quantifying uncertainty in the biospheric carbon flux for England and Wales. *J. Roy. Statist. Soc. Ser. A* **171** 109–135. MR2412649

LEE, L. A., CARSLAW, K. S., PRINGLE, K. J., MANN, G. W. and SPRACKLEN, D. V. (2011). Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters. *Atmos. Chem. Phys.* **11** 12253–12273.

LEE, L. A., CARSLAW, K. S., PRINGLE, K. J. and MANN, G. W. (2012). Mapping the uncertainty in global CCN using emulation. *Atmospheric Chemistry and Physics* **12** 9739–9751.

LI, R. and SUDJIANTO, A. (2005). Analysis of computer experiments using penalized likelihood in Gaussian kriging models. *Technometrics* **47** 111–120. MR2188073

LINDSAY, B. G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes* (*Ithaca*, *NY*, 1987). *Contemp. Math.* **80** 221–239. Amer. Math. Soc., Providence, RI. MR0999014

LINDSAY, B. G., YI, G. Y. and SUN, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statist. Sinica* **21** 71–105. MR2796854

LINKLETTER, C., BINGHAM, D., HENGARTNER, N., HIGDON, D. and YE, K. Q. (2006). Variable selection for Gaussian process models in computer experiments. *Technometrics* **48** 478–490. MR2328617

LOPES, D. (2011). Development and implementation of Bayesian computer model emulators. Ph.D. thesis, Duke Univ.

MARREL, A., IOOSS, B., JULLIEN, M., LAURENT, B. and VOLKOVA, E. (2011). Global sensitivity analysis for models with spatially dependent outputs. *Environmetrics* **22** 383–397. MR2843392

PATRA, A. K., BAUER, A. C., NICHITA, C. C., PITMAN, E. B., SHERIDAN, M. F., BURSIK, M., RUPP, B., WEBBER, A., STINTON, A. J., NAMIKAWA, L. M. et al. (2005). Parallel adaptive numerical simulation of dry avalanches over natural terrain. *J. Volcanol. Geotherm. Res.* **139** 1–21.

PAULO, R. (2005). Default priors for Gaussian processes. *Ann. Statist.* **33** 556–582. MR2163152

PAULO, R., GARCÍA-DONATO, G. and PALOMO, J. (2012). Calibration of computer models with multivariate output. *Comput. Statist. Data Anal.* **56** 3959–3974. MR2957846

PITMAN, E. B., NICHITA, C. C., PATRA, A., BAUER, A., SHERIDAN, M. and BURSIK, M. (2003). Computing granular avalanches and landslides. *Phys. Fluids* **15** 3638–3646. MR2028451

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

REN, C., SUN, D. and HE, C. (2012). Objective Bayesian analysis for a spatial model with nugget effects. *J. Statist. Plann. Inference* **142** 1933–1946. MR2903403

ROUGIER, J. (2008). Efficient emulators for multivariate deterministic functions. *J. Comput. Graph. Statist.* **17** 827–843. MR2649069

ROUGIER, J., GUILLAS, S., MAUTE, A. and RICHMOND, A. D. (2009). Expert knowledge and multivariate emulation: The thermosphere-ionosphere electrodynamics general circulation model (TIE-GCM). *Technometrics* **51** 414–424. MR2756477

ROUSTANT, O., GINSBOURGER, D. and DEVILLE, Y. (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *J. Stat. Softw.* **51** 1–55.

SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. MR1041765

SAVITSKY, T., VANNUCCI, M. and SHA, N. (2011). Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statist. Sci.* **26** 130–149. MR2849913

SEVERINI, T. A. (2000). *Likelihood Methods in Statistics*. *Oxford Statistical Science Series* **22**. Oxford Univ. Press, Oxford. MR1854870

SPILLER, E. T., BAYARRI, M. J., BERGER, J. O., CALDER, E. S., PATRA, A. K., PITMAN, E. B. and WOLPERT, R. L. (2014). Automating emulator construction for geophysical hazard maps. *SIAM/ASA J. Uncertain. Quantificat.* **2** 126–152. MR3283903

VARIN, C., REID, N. and FIRTH, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42. MR2796852

XIAO, M., BREITKOPF, P., FILOMENO COELHO, R., KNOPF-LENOIR, C., SIDORKIEWICZ, M. and VILLON, P. (2010). Model reduction by CPOD and Kriging: Application to the shape optimization of an intake port. *Struct. Multidiscip. Optim.* **41** 555–574. MR2601473

ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261. MR2054303

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1348–1372
DOI: 10.1214/16-AOAS938

A HIERARCHICAL FRAMEWORK FOR STATE-SPACE MATRIX INFERENCE AND CLUSTERING

BY CHANDLER ZUO, KAILEI CHEN, KYLE J. HEWITT, EMERY H. BRESNICK AND SÜNDÜZ KELEŞ[1]

University of Wisconsin–Madison

Integrative analysis of multiple experimental datasets measured over a large number of observational units is the focus of large numbers of contemporary genomic and epigenomic studies. The key objectives of such studies include not only inferring a hidden state of activity for each unit over individual experiments, but also detecting highly associated clusters of units based on their inferred states. Although there are a number of methods tailored for specific datasets, there is currently no state-of-the-art modeling framework for this general class of problems. In this paper, we develop the MBASIC (*M*atrix *B*ased *A*nalysis for *S*tate-space *I*nference and *C*lustering) framework. MBASIC consists of two parts: state-space mapping and state-space clustering. In state-space mapping, it maps observations onto a finite state-space, representing the activation states of units across conditions. In state-space clustering, MBASIC incorporates a finite mixture model to cluster the units based on their inferred state-space profiles across all conditions. Both the state-space mapping and clustering can be simultaneously estimated through an Expectation-Maximization algorithm. MBASIC flexibly adapts to a large number of parametric distributions for the observed data, as well as the heterogeneity in replicate experiments. It allows for imposing structural assumptions on each cluster, and enables model selection using information criterion. In our data-driven simulation studies, MBASIC showed significant accuracy in recovering both the underlying state-space variables and clustering structures. We applied MBASIC to two genome research problems using large numbers of datasets from the ENCODE project. The first application grouped genes based on transcription factor occupancy profiles of their promoter regions in two different cell types. The second application focused on identifying groups of loci that are similar to a GATA2 binding site that is functional at its endogenous locus by utilizing transcription factor occupancy data and illustrated applicability of MBASIC in a wide variety of problems. In both studies, MBASIC showed higher levels of raw data fidelity than analyzing these data with a two-step approach using ENCODE results on transcription factor occupancy data.

REFERENCES

ANANDAPADAMANABAN, M., ANDRESEN, C., HELANDER, S., OHYAMA, Y., SIPONEN, M. I., LUNDSTRÖM, P., KOKUBO, T., IKURA, M., MOCHE, M. and SUNNERHAGEN, M. (2013). High-resolution structure of TBP with TAF1 reveals anchoring patterns in transcriptional regulation. *Nat. Struct. Mol. Biol.* **20** 1008–1014.

Key words and phrases. State-space, clustering, E-M algorithm, transcription factors, ChIP-seq.

ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*. **11** R106.

CHENG, C., YAN, K.-K., HWANG, W., QIAN, J., BHARDWAJ, N., ROZOWSKY, J., LU, Z. J., NIU, W., ALVES, P., KATO, M., SNYDER, M. and GERSTEIN, M. (2011). Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput*. *Biol*. **7**.

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

DORÉ, L. C., CHLON, T. M., BROWN, C. D., WHITE, K. P. and CRISPINO, J. D. (2012). Chromatin occupancy analysis reveals genome-wide GATA factor switching during hematopoiesis. *Blood* **119** 3724–3733.

ENCODE PROJECT CONSORTIUM (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74.

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist*. *Assoc*. **97** 611–631. MR1951635

GAO, X., JOHNSON, K. D., CHANG, Y.-I., BOYER, M. E., DEWEY, C. N., ZHANG, J. and BRESNICK, E. H. (2013). Gata2 cis-element is required for hematopoietic stem cell generation in the mammalian embryo. *J. Exp*. *Med*. **210** 2833–2842.

GERSTEIN, M. B., KUNDAJE, A., HARIHARAN, M., LANDT, S. G., YAN, K.-K., CHENG, C., MU, X. J., KHURANA, E., ROZOWSKY, J., ALEXANDER, R., MIN, R., ALVES, P., ABYZOV, A., ADDLEMAN, N., BHARDWAJ, N., BOYLE, A. P., CAYTING, P., CHAROS, A., CHEN, D. Z., CHENG, Y., CLARKE, D., EASTMAN, C., EUSKIRCHEN, G., FRIETZE, S., FU, Y., GERTZ, J., GRUBERT, F., HARMANCI, A., JAIN, P., KASOWSKI, M., LACROUTE, P., LENG, J., LIAN, J., MONAHAN, H., O'GEEN, H., OUYANG, Z., PARTRIDGE, E. C., PATACSIL, D., PAULI, F., RAHA, D., RAMIREZ, L., REDDY, T. E., REED, B., SHI, M., SLIFER, T., WANG, J., WU, L., YANG, X., YIP, K. Y., ZILBERMAN-SCHAPIRA, G., BATZOGLOU, S., SIDOW, A., FARNHAM, P. J., MYERS, R. M., WEISSMAN, S. M. and SNYDER, M. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature* **489** 91–100.

HOLLEY, D. W., GROH, B. S., WOZNIAK, G., DONOHOE, D. R., SUN, W., GODFREY, V. and BULTMAN, S. J. (2014). The BRG1 chromatin remodeler regulates widespread changes in gene expression and cell proliferation during B cell activation. *J. Cell*. *Physiol*. **229** 44–52.

HSU, A. P., JOHNSON, K. D., FALCONE, E. L., SANALKUMAR, R., SANCHEZ, L., HICKSTEIN, D. D., CUELLAR-RODRIGUEZ, J., LEMIEUX, J. E., ZERBE, C. S., BRESNICK, E. H. and HOLLAND, S. M. (2013). GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome. *Blood* **121** 3830–3837.

HU, G., SCHONES, D. E., CUI, K., YBARRA, R., NORTHRUP, D., TANG, Q., GATTINONI, L., RESTIFO, N. P., HUANG, S. and ZHAO, K. (2011). Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res*. **21** 1650–1658.

JI, H., LI, X., WANG, Q. and NING, Y. (2013). Differential principle component analysis of ChIP-seq. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **110** 6789–6794.

JOHNSON, K. D., HSU, A. P., RYU, M.-J., WANG, J., GAO, X., BOYER, M. E., LIU, Y., LEE, Y., CALVO, K. R., KELES, S., ZHANG, J., HOLLAND, S. M. and BRESNICK, E. H. (2012). Cis-element mutation in a GATA-2-dependent immunodeficiency syndrome governs hematopoiesis and vascular integrity. *J. Clin*. *Invest*. **10** 3692–3704.

KIM, S.-I., BRESNICK, E. H. and BULTMAN, S. J. (2009). BRG1 directly regulates nucleosome structure and chromatin looping of the α globin locus to activate transcription. *Nucleic Acids Res*. **37** 6019–6027.

KIM, S.-I., BULTMAN, S. J., KIEFER, C. M., DEAN, A. and BRESNICK, E. H. (2009). BRG1 requirement for long-range interaction of a locus control region with a downstream promoter. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **106** 2259–2264.

KUNARSO, G., CHIA, N.-Y., JEYAKANI, J., HWANG, C., LU, X., CHAN, Y.-S., NG, H.-H. and BOURQUE, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42** 631–634.

LEE, S., HUANG, J. Z. and HU, J. (2010). Sparse logistic principal components analysis for binary data. *Ann. Appl. Stat.* **4** 1579–1601. MR2758342

LIANG, K. and KELES, S. (2012). Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics* **28** 121–122.

LINNEMAN, A. K., O'GEEN, H., KELEŞ, S., FARNHAM, P. J. and BRESNICK, E. H. (2011). Genetic framework for GATA factor function in vascular biology. *Proc. Natl. Acad. Sci. USA* **108** 13641–13646.

NEPH, S., STERGACHIS, A. B., REYNOLDS, A., SANDSTROM, R., BORENSTEIN, E. and STAMATOYANNOPOULOS, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150** 1274–1286.

ROY, S., WAPINSKI, I., PFIFFNER, J., FRENCH, C., SOCHA, A., KONIECZKA, J., HABIB, N., KELLIS, M., THOMPSON, D. and REGEV, A. (2013). Arboretum: Reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Res.* **23** 1039–1050.

SCHMIDT, D., WILSON, M. D., BALLESTER, B., SCHWALIE, P. C., BROWN, G. D., MARSHALL, A., KUTTER, C., WATT, S., MARTINEZ-JIMENEZ, C. P., MACKAY, S., TALIANIDIS, I., FLICEK, P. and ODOM, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328** 1036–1040.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. and MESIROV, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102** 15545–15550.

WALTMAN, P., KACMARCZYK, T., BATE, A. R., KEARNS, D. B., REISS, D. J., EICHENBERGER, P. and BONNEAU, R. (2010). Multi-species integrative biclustering. *Genome Biol.* **11** R96.

WANG, J., ZHUANG, J., IYER, S., LIN, X., WHITFIELD, T. W., GREVEN, M. C., PIERCE, B. G., DONG, X., KUNDAJE, A., CHENG, Y., RANDO, O. J., BIRNEY, E., MYERS, R. M., NOBLE, W. S., SNYDER, M. and WENG, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22** 1798–1812.

WEI, Y., TENZEN, T. and JI, H. (2015). Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics* **16** 31–46. MR3365409

WEI, Y., LI, X., WANG, Q. and JI, H. (2012). iaseq: Integrative analysis of allele-specificity of protein-dna interactions in multiple chip-seq datasets. *BMC Genomics* **13** 1–19.

ZENG, X., SANALKUMAR, R., BRESNICK, E. H., LI, H., CHANG, Q. and KELEŞ, S. (2013). jMOSAiCS: Joint analysis of multiple ChIP-seq datasets. *Genome Biol.* **14** R38.

ZUO, C., CHEN, K., HEWITT, K. J., BRESNICK, E. H. and KELEŞ, S. (2016). Supplement to "A hierarchical framework for state-space matrix inference and clustering." DOI:10.1214/16-AOAS938SUPP.

ZUO, C. and KELEŞ, S. (2014). A statistical framework for power calculations in ChIP-seq experiments. *Bioinformatics* **30** 753–760.

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1373–1396
DOI: 10.1214/16-AOAS939

DETECTION OF EPIGENOMIC NETWORK COMMUNITY ONCOMARKERS

BY THOMAS E. BARTLETT[1] AND ALEXEY ZAIKIN[2]

University College London

In this paper we propose network methodology to infer prognostic cancer biomarkers based on the epigenetic pattern DNA methylation. Epigenetic processes such as DNA methylation reflect environmental risk factors, and are increasingly recognised for their fundamental role in diseases such as cancer. DNA methylation is a gene-regulatory pattern, and hence provides a means by which to assess genomic regulatory interactions. Network models are a natural way to represent and analyse groups of such interactions. The utility of network models also increases as the quantity of data and number of variables increase, making them increasingly relevant to large-scale genomic studies. We propose methodology to infer prognostic genomic networks from a DNA methylation-based measure of genomic interaction and association. We then show how to identify prognostic biomarkers from such networks, which we term "network community oncomarkers". We illustrate the power of our proposed methodology in the context of a large publicly available breast cancer dataset.

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.

BARABÁSI, A.-L. and OLTVAI, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **5** 101–113.

BARTLETT, T. E. (2015). Network inference and community detection, based on covariance matrices, correlations and test statistics from arbitrary distributions. Preprint. Available at arXiv:1506.04928.

BARTLETT, T. E., OLHEDE, S. C. and ZAIKIN, A. (2014). A DNA methylation network interaction measure, and detection of network oncomarkers. *PLoS ONE* **9** e84573.

BARTLETT, T. E. and ZAIKIN, A. (2016). Supplement to "Detection of epigenomic network community oncomarkers." DOI:10.1214/16-AOAS939SUPP.

BARTLETT, T. E., ZAIKIN, A., OLHEDE, S. C., WEST, J., TESCHENDORFF, A. E. and WIDSCHWENDTER, M. (2013). Corruption of the intra-gene DNA methylation architecture is a hallmark of cancer. *PLoS ONE* **8** e68285.

BEGUERISSE-DÍAZ, M., GARDUÑO-HERNÁNDEZ, G., VANGELOV, B., YALIRAKI, S. N. and BARAHONA, M. (2014). Interest communities and flow roles in directed networks: The Twitter network of the UK riots. *J. R. Soc. Interface* **11** 20140940.

BHAGAT, R., CHADAGA, S., PREMALATA, C. S., RAMESH, G., RAMESH, C., PALLAVI, V. R. and KRISHNAMOORTHY, L. (2012). Aberrant promoter methylation of the RASSF1A and APC genes in epithelial ovarian carcinoma development. *Cellular Oncology* **35** 473–479.

Key words and phrases. Computational biology, stochastic networks, community detection, epigenomics.

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.
BONETTA, L. (2006). Genome sequencing in the fast Lane. *Nature Methods* **3** 141.
BROCKS, D., ASSENOV, Y., MINNER, S., BOGATYROVA, O., SIMON, R., KOOP, C., OAKES, C., ZUCKNICK, M., LIPKA, D. B., WEISCHENFELDT, J. et al. (2014). Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Reports* **8** 798–806.
CHRISTENSEN, B. C., HOUSEMAN, E. A., MARSIT, C. J., ZHENG, S., WRENSCH, M. R., WIEMELS, J. L., NELSON, H. H., KARAGAS, M. R., PADBURY, J. F., BUENO, R. et al. (2009). Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genetics* **5** e1000602.
CLUNE, J., MOURET, J.-B. and LIPSON, H. (2013). The evolutionary origins of modularity. *Proc. R. Soc. Lond., B Biol. Sci.* **280** 20122863.
COLLINS, F. and BARKER, A. (2007). Mapping the cancer genome. *Scientific American Magazine* **296** 50–57.
COONEY, C. A. (2007). Epigenetics–DNA-based mirror of our environment? *Dis. Markers* **23** 121–137.
COX, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **34** 187–220. MR0341758
FEINBERG, A. P., OHLSSON, R. and HENIKOFF, S. (2006). The epigenetic progenitor origin of human cancer. *Nat. Rev. Genet.* **7** 21–33.
FLEISCHER, T., FRIGESSI, A., JOHNSON, K. C., EDVARDSEN, H., TOULEIMAT, N., KLAJIC, J., RIIS, M. L., HAAKENSEN, V., WÄRNBERG, F., NAUME, B. et al. (2014). Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome Biol* **15** 435.
GAO, F., SHI, L., RUSSIN, J., ZENG, L., CHANG, X., HE, S., CHEN, T. C., GIANNOTTA, S. L., WEISENBERGER, D. J., ZADA, G. et al. (2013). DNA methylation in the malignant transformation of meningiomas. *PloS One* **8** e54114.
GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826 (electronic). MR1908073
HAMPTON, T. (2006). Cancer genome atlas. *JAMA*: *The Journal of the American Medical Association* **296** 1958–1958.
HARRELL, F. E. (2001). *Regression Modeling Strategies*: *With Applications to Linear Models*, *Logistic Regression*, *and Survival Analysis*. Springer, Berlin.
HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137. MR0718088
HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika* **28** 321–377.
JACOB, L., NEUVIAL, P. and DUDOIT, S. (2012). More power via graph-structured tests for differential expression of gene networks. *Ann. Appl. Stat.* **6** 561–600. MR2976483
JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135
JONES, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13** 484–492.
KANG, G. H., SHIM, Y.-H., JUNG, H.-Y., KIM, W. H., RO, J. Y. and RHYU, M.-G. (2001). CpG island methylation in premalignant stages of gastric carcinoma. *Cancer Research* **61** 2847–2851.
KANG, G. H., LEE, S., KIM, J.-S. and JUNG, H.-Y. (2003). Profile of aberrant CpG island methylation along multistep gastric carcinogenesis. *Laboratory Investigation* **83** 519–526.
KATENKA, N. and KOLACZYK, E. D. (2012). Inference and characterization of multi-attribute networks with application to computational biology. *Ann. Appl. Stat.* **6** 1068–1094. MR3012521
KISHIDA, Y., NATSUME, A., KONDO, Y., TAKEUCHI, I., AN, B., OKAMOTO, Y., SHINJO, K., SAITO, K., ANDO, H., OHKA, F. et al. (2012). Epigenetic subclassification of meningiomas based on genome-wide DNA methylation analyses. *Carcinogenesis* **33** 436–441.

LAI, F. and SHIEKHATTAR, R. (2014). Where long noncoding RNAs meet DNA methylation. *Cell Res*. **24** 263–264.

LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2011). Overlapping stochastic block models with application to the French political blogosphere. *Ann. Appl. Stat*. **5** 309–336. MR2810399

LI, C. and LI, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl. Stat*. **4** 1498–1516. MR2758338

LI, C. and WANG, J. (2014). Quantifying the underlying landscape and paths of cancer. *J. R. Soc. Interface* **11** 20140774.

LUO, Y., WONG, C.-J., KAZ, A. M., DZIECIATKOWSKI, S., CARTER, K. T., MORRIS, S. M., WANG, J., WILLIS, J. E., MAKAR, K. W., ULRICH, C. M. et al. (2014). Differences in DNA methylation signatures reveal multiple pathways of progression from adenoma to colorectal cancer. *Gastroenterology* **147** 418–429.

MAEKAWA, R., SATO, S., YAMAGATA, Y., ASADA, H., TAMURA, I., LEE, L., OKADA, M., TAMURA, H., TAKAKI, E., NAKAI, A. et al. (2013). Genome-wide DNA methylation analysis reveals a potential mechanism for the pathogenesis and development of uterine leiomyomas. *PloS One* **8** e66632.

MARDIA, K. V. (2013). Statistical approaches to three key challenges in protein structural bioinformatics. *J. R. Stat. Soc. Ser. C. Appl. Stat*. **62** 487–514. MR3060628

NANDI, A. K., SUMANA, A. and BHATTACHARYA, K. (2014). Social insect colony as a biological regulatory system: Modelling information flow in dominance networks. *J. R. Soc. Interface* **11** 20140951.

NAVARRO, A., YIN, P., MONSIVAIS, D., LIN, S. M., DU, P., WEI, J.-J. and BULUN, S. E. (2012). Genome-wide DNA methylation indicates silencing of tumor suppressor genes in uterine leiomyoma. *PLoS ONE* **7** e33284.

NEWMAN, M. E. (2004). Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems* **38** 321–330.

NEWMAN, M. E. and GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* (3) **69** 026113.

OLHEDE, S. C. and WOLFE, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proc. Natl. Acad. Sci. USA* **111** 14722–14727.

PALLA, G., LOVÁSZ, L. and VICSEK, T. (2010). Multifractal network generator. *Proc. Natl. Acad. Sci. USA* **107** 7640–7645.

PENG, J., ZHU, J., BERGAMASCHI, A., HAN, W., NOH, D.-Y., POLLACK, J. R. and WANG, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat*. **4** 53–77. MR2758084

QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems* 3120–3128. Lake Tahoe, Nevada.

REZNIK, E., WATSON, A. and CHAUDHARY, O. (2013). The stubborn roots of metabolic cycles. *J. R. Soc. Interface* **10** 20130087.

RIOLO, M. A. and NEWMAN, M. E. J. (2012). First-principles multiway spectral partitioning of graphs. Preprint. Available at arXiv:1209.5969.

SAAVEDRA, S., ROHR, R. P., GILARRANZ, L. J. and BASCOMPTE, J. (2014). How structurally stable are global socioeconomic systems? *J. R. Soc. Interface* **11** 20140693.

SHEN-ORR, S. S., MILO, R., MANGAN, S. and ALON, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet*. **31** 64–68.

TAYLOR, I. W., LINDING, R., WARDE-FARLEY, D., LIU, Y., PESQUITA, C., FARIA, D., BULL, S., PAWSON, T., MORRIS, Q. and WRANA, J. L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol*. **27** 199–204.

TRAN, T.-D. and KWON, Y.-K. (2013). The relationship between modularity and robustness in signalling networks. *J. R. Soc. Interface* **10** 20130771.

VAN HOESEL, A. Q., SATO, Y., ELASHOFF, D. A., TURNER, R. R., GIULIANO, A. E., SHAMONKI, J. M., KUPPEN, P. J. K., VAN DE VELDE, C. J. H. and HOON, D. S. B. (2013). Assessment of DNA methylation status in early stages of breast cancer development. *British Journal of Cancer* **108** 2033–2038.

VENTERS, B. J. and PUGH, B. F. (2013). Genomic organization of human transcription initiation complexes. *Nature* **502** 53–58.

VERSCHUUR-MAES, A. H., DE BRUIN, P. C. and VAN DIEST, P. J. (2012). Epigenetic progression of columnar cell lesions of the breast to invasive breast cancer. *Breast Cancer Res. Treat.* **136** 705–715.

VU, D. Q., HUNTER, D. R. and SCHWEINBERGER, M. (2013). Model-based clustering of large networks. *Ann. Appl. Stat.* **7** 1010–1039. MR3113499

WAGNER, A. (2002). Estimating coarse gene network structure from large-scale gene perturbation data. *Genome Res.* **12** 309–315.

WEI, P. and PAN, W. (2010). Network-based genomic discovery: Application and comparison of Markov random-field models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59** 105–125. MR2750134

XIE, W., SCHULTZ, M. D., LISTER, R., HOU, Z., RAJAGOPAL, N., RAY, P., WHITAKER, J. W., TIAN, S., HAWKINS, R. D., LEUNG, D. et al. (2013). Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153** 1134–1148.

YAMAMOTO, E., SUZUKI, H., YAMANO, H., MARUYAMA, R., NOJIMA, M., KAMIMAE, S., SAWADA, T., ASHIDA, M., YOSHIKAWA, K., KIMURA, T. et al. (2012). Molecular dissection of premalignant colorectal lesions reveals early onset of the CpG island methylator phenotype. *Am. J. Pathol.* **181** 1847–1861.

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1397–1426
DOI: 10.1214/16-AOAS940

SPARSE MEDIAN GRAPHS ESTIMATION IN A HIGH-DIMENSIONAL SEMIPARAMETRIC MODEL

BY FANG HAN*, XIAOYAN HAN†, HAN LIU† AND BRIAN CAFFO*

*Johns Hopkins University** *and Princeton University*†

We propose a unified framework for conducting inference on complex aggregated data in high-dimensional settings. We assume the data are a collection of multiple non-Gaussian realizations with underlying undirected graphical structures. Using the concept of median graphs in summarizing the commonality across these graphical structures, we provide a novel semiparametric approach to modeling such complex aggregated data, along with robust estimation of the median graph, which is assumed to be sparse. We prove the estimator is consistent in graph recovery and give an upper bound on the rate of convergence. We further provide thorough numerical analysis on both synthetic and real datasets to illustrate the empirical usefulness of the proposed models and methods.

REFERENCES

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

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*. Springer, Heidelberg. MR2807761

BULLMORE, E. and SPORNS, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10** 186–198.

BUNKE, H. and SHEARER, K. (1998). A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters* **19** 255–259.

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

DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175.

ELOYAN, A., MUSCHELLI, J., NEBEL, M. B., LIU, H., HAN, F., ZHAO, T., BARBER, A., JOEL, S., PEKAR, J. J., MOSTOFSKY, S. and CAFFO, B. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience* **6** 61.

FINGELKURTS, A. A. and KÄHKÖNEN, S. (2005). Functional connectivity in the brain—Is it an elusive concept? *Neuroscience and Biobehavioral Reviews* **28** 827–836.

FRIEDMAN, J. H., HASTIE, T. and TIBSHIRANI, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

FRISTON, K. J. (2011). Functional and effective connectivity: A review. *Brain Connect.* **1** 13–36.

HAN, F. and LIU, H. (2014). Distribution-free tests of independence with applications to testing more structures. Preprint. Available at arXiv:1410.4179.

HORWITZ, B. (2003). The elusive concept of brain connectivity. *Neuroimage* **19** 466–470.

Key words and phrases. Graphical model, median graph, complex aggregated data, semiparametric model, high-dimensional statistics.

HSIEH, C. J., SUSTIK, M. A., RAVIKUMAR, P. and DHILLON, I. S. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems* (*NIPS*) **24**. Granada, Spain.

JIANG, X., MUNGER, A. and BUNKE, H. (2001). On median graphs: Properties, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23** 1144–1151.

JIN, J., ZHANG, C.-H. and ZHANG, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *J. Mach. Learn. Res.* **15** 2723–2772. MR3270749

KE, T., JIN, J. and FAN, J. (2014). Covariance assisted screening and estimation. *Ann. Statist.* **42** 2202.

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

LEDOIT, O. and WOLF, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* **10** 603–621.

LI, L. and TOH, K.-C. (2010). An inexact interior point method for L_1-regularized sparse covariance selection. *Math. Program. Comput.* **2** 291–315. MR2741488

LIU, I. and AGRESTI, A. (1996). Mantel–Haenszel-type infererence for cumulative odds ratios with a stratified ordinal response. *Biometrics* **52** 1223–1234. MR1422076

LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

LIU, W. and LUO, X. (2012). High-dimensional sparse precision matrix estimation via sparse column inverse operator. Preprint. Available at arXiv:1203.3896.

LIU, H., ROEDER, K. and WASSERMAN, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. In *Advances in Neural Information Processing Systems* 1432–1440. Vancouver, Canada.

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

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

MILHAM, M. P., FAIR, D., MENNES, M. and MOSTOFSKY, S. H. (2012). The ADHD-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience* **6** 62.

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

POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M., SCHLAGGAR, B. L. and PETERSON, S. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.

RAMSAY, J. D., HANSON, S. J., HANSON, C., HALCHENKO, Y., POLDRACK, R. and GLYMOUR, C. (2009). Six problems for causal inference from fMRI. *NeuroImage* **49** 1545–1558.

RAVIKUMAR, P., WAINWRIGHT, M., RASKUTTI, G. and YU, B. (2009). Model selection in Gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. In *Advances in Neural Information Processing Systems* **22**. Vancouver, Canada.

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

ROWE, B. L. Y. (2014). tawny: Provides various portfolio optimization strategies including random matrix theory and shrinkage estimators. R package version 2.1.2.

RUBINOV, M. and SPORNS, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *Neuroimage* **52** 1059–1069.

SCHEINBERG, K., MA, S. and GLODFARB, D. (2010). Sparse inverse covariance selection via alternating linearization methods. In *Advances in Neural Information Processing Systems* (*NIPS*) **23**. Vancouver, Canada.

XU, W., HOU, Y., HUNG, Y. S. and ZOU, Y. (2010). Comparison of Spearman's rho and Kendall's tau in normal and contaminated normal models. Preprint. Available at arXiv:1011.2009.

XUE, L. and ZOU, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.* **40** 2541–2571. MR3097612

YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

ZHAO, T., LIU, H., ROEDER, K., LAFFERTY, J. and WASSERMAN, L. (2012). The `huge` package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13** 1059–1062. MR2930633

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1427–1446
DOI: 10.1214/16-AOAS941

QUANTIFYING THE SPATIAL INEQUALITY AND TEMPORAL TRENDS IN MATERNAL SMOKING RATES IN GLASGOW[1]

BY DUNCAN LEE AND ANDREW LAWSON

University of Glasgow and Medical University of South Carolina

Maternal smoking is well known to adversely affect birth outcomes, and there is considerable spatial variation in the rates of maternal smoking in the city of Glasgow, Scotland. This spatial variation is a partial driver of health inequalities between rich and poor communities, and it is of interest to determine the extent to which these inequalities have changed over time. Therefore in this paper we develop a Bayesian hierarchical model for estimating the spatio-temporal pattern in smoking incidence across Glasgow between 2000 and 2013, which can identify the changing geographical extent of clusters of areas exhibiting elevated maternal smoking incidences that partially drive health inequalities. Additionally, we provide freely available software via the R package `CARBayesST` to allow others to implement the model we have developed. The study period includes the introduction of a ban on smoking in public places in 2006, and the results show an average decline of around 11% in maternal smoking rates over the study period.

REFERENCES

ANDERSON, C., LEE, D. and DEAN, N. (2014). Identifying clusters in Bayesian disease mapping. *Biostat*. **15** 457–469.

BAULD, L., FERGUSON, J., LAWSON, L., CHESTERMAN, J. and JUDGE, K. (2005). Tackling smoking in Glasgow. Technical report, Glasgow Centre for Population Health.

CHARRAS-GARRIDO, M., ABRIAL, D. and DE GOER, J. (2012). Classification method for disease risk mapping based on discrete hidden Markov random fields. *Biostat*. **13** 241–255.

CHARRAS-GARRIDO, M., AZIZI, L., FORBES, F., DOYLE, S., PEYRARD, N. and ABRIAL, D. (2013). On the difficulty to delimit disease risk hot sports. *Journal of Applied Earth Observation and Geoinformation* **22** 99–105.

CHOI, J., LAWSON, A. B., CAI, B. and HOSSAIN, MD. M. (2011). Evaluation of Bayesian spatiotemporal latent models in small area health data. *Environmetrics* **22** 1008–1022. MR2861574

CNATTINGIUS, S. (2004). The epidemiology of smoking during pregnancy: Smoking prevalence, maternal characteristics, and pregnancy outcomes. *Nicotine Tob. Res*. **6 Suppl 2** S125–S140.

EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci*. **11** 89–121. MR1435485

FORBES, F., CHARRAS-GARRIDO, M., AZIZI, L., DOYLE, S. and ABRIAL, D. (2013). Spatial risk mapping for rare disease with hidden Markov fields and variational EM. *Ann. Appl. Stat*. **7** 1192–1216. MR3113506

GANGNON, R. and CLAYTON, M. (2000). Bayesian detection and modeling of spatial disease clustering. *Biometrics* **56** 922–935.

Key words and phrases. Cluster detection, maternal smoking, spatial inequality, spatio-temporal modelling.

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

GRAY, L. and LEYLAND, A. (2009). "Glasgow effect" of cigarette smoking explained by socio-economic status? A multilevel analysis. *BMC Public Health* **9** 245.

GRAY, R., BONELLIE, S., CHALMERS, J., GREER, I., JARVIS, S., KURINCZUK, J. and WILLIAMS, C. (2009). Contribution of smoking during pregnancy to inequalities in stillbirth and infant death in Scotland 1994–2003: Retrospective population based study using hospital maternity records. *British Medical Journal* **339** b3754.

GRAY, L., MERLO, J., MINDELL, J., HALLQVIST, J., TAFFOREAU, J., O'REILLY, D., REGIDOR, E., NAESS, O., KELLEHER, C., HELAKORPI, S., LANGE, C. and LEYLAND, A. (2012). International differences in self-reported health measures in 33 major metropolitan areas in Europe. *European Journal of Public Health* **22** 40–47.

GREEN, P. J. and RICHARDSON, S. (2002). Hidden Markov models and disease mapping. *J. Amer. Statist. Assoc.* **97** 1055–1070. MR1951259

KNORR-HELD, L. (2000). Bayesian modelling of inseparable space–time variation in disease risk. *Stat. Med.* **19** 2555–2567.

KNORR-HELD, L. and RASSER, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56** 13–21.

KRAMER, M., SEGUIN, L., LYDON, J. and GOULET, L. (2000). Socio-economic disparities in pregnancy outcome: Why do the poor fare so poorly? *Paediatr. Perinat. Epidemiol.* **14** 194–210.

KULLDORFF, M., HEFFERNAN, R., HARTMAN, J., ASSUNCAO, R. and MOSTASHARI, F. (2005). A space-time permutation scan statistic for disease outbreak detection. *PLoS Med.* **2** 216–224.

LAWSON, A. B. (2009). *Bayesian Disease Mapping*: *Hierarchical Modeling in Spatial Epidemiology*. CRC Press, Boca Raton, FL. MR2484272

LAWSON, A. B., CHOI, J., CAI, B., HOSSAIN, M., KIRBY, R. S. and LIU, J. (2012). Bayesian 2-stage space-time mixture modeling with spatial misalignment of the exposure in small area health data. *J. Agric. Biol. Environ. Stat.* **17** 417–441. MR2993274

LEE, D. and LAWSON, A. (2016). Supplement to "Quantifying the spatial inequality and temporal trends in maternal smoking rates in Glasgow." DOI:10.1214/16-AOAS941SUPPA, DOI:10.1214/16-AOAS941SUPPB.

LEROUX, B. G., LEI, X. and BRESLOW, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (*Minneapolis, MN*, 1997). *IMA Vol. Math. Appl.* **116** 179–191. Springer, New York. MR1731684

LI, G., BEST, N., HANSELL, A., AHMED, I. and RICHARDSON, S. (2012). BaySTDetect: Detecting unusual temporal patterns in small area data via Bayesian model choice. *Biostat.* **13** 695–710.

MACKAY, D., NELSON, S., HAW, S. and PELL, J. (2012). Impact of Scotland's smoke-free llegislation on pregnancy complications: Retrospective cohort study. *PLoS Med.* **9** e1001175.

PASSMORE, E., MCGUIRE, R., CORRELL, P. and BENTLEY, J. (2015). Demographic factors associated with smoking cessation during pregnancy in New South Wales, Australia, 2000–2011. *BMC Public Health* **15** 398.

RAND, W. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.

RUSHWORTH, A., LEE, D. and MITCHELL, R. (2014). A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial and Spatio-temporal Epidemiology* **10** 29–38.

SILVEIRA, M., MATIJASEVICH, A., MENEZES, A., HORTA, B., SANTOS, I., DARROS, A., BARROS, F. and VICTORA, C. (2016). Secular trends in smoking during pregnancy according to income and ethnic group: Four population-based perinatal surveys in a Brazilian city. *BMJ Open* **6** e010127.

TAPPIN, D., MACASKILL, S., BAULD, L., EADIE, D., SHIPTON, D. and GALBRAITH, L. (2010). Smoking prevalence and smoking cessation services for pregnant women in Scotland. *Substance Abuse Treatment, Prevention and Policy* **5** 1.

R CORE TEAM (2013). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

UGARTE, D., ETXEBERRIA, J., GOICOA, T. and ARDANAZ, E. (2012). Gender-specific spatio-temporal patterns of colorectal cancer incidence in Navarre, Spain (1990–2005). *Cancer Epidemiol*. **36** 254–262.

WAKEFIELD, J. and KIM, A. (2013). A Bayesian model for cluster detection. *Biostat*. **14** 752–765.

WANG, X., ZUCKERMAN, B., PEARSON, C., KAUFMAN, G., CHEN, C., WANG, G., NIU, T., WISE, P., BAUCHNER, H. and XU, X. (2002). Maternal cigarette smoking, metabolic gene polymorphism, and infant birth weight. *J. Am. Med. Assoc*. **287** 195–202.

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

WILLIAMSON, D., SERDULA, M., KENDRICK, J. and BINKIN, N. (1989). Comparing the prevalence of smoking in pregnant and nonpregnant women. *J. Am. Med. Assoc*. **261** 70–74.

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1447–1471
DOI: 10.1214/16-AOAS942

USING SCHEFFÉ PROJECTIONS FOR MULTIPLE OUTCOMES IN AN OBSERVATIONAL STUDY OF SMOKING AND PERIODONTAL DISEASE

BY PAUL R. ROSENBAUM

University of Pennsylvania

In an observational study of the effects caused by treatments, a sensitivity analysis asks about the magnitude of bias from unmeasured covariates that would need to be present to alter the conclusions of a naive analysis that presumes adjustments for measured covariates remove all biases. When there are two or more outcomes in an observational study, these outcomes may be unequally sensitive to unmeasured biases, and the least sensitive finding may concern a combination of several outcomes. A method of sensitivity analysis is proposed using Scheffé projections that permits the investigator to consider all linear contrasts in two or more scored outcomes while controlling the family-wise error rate. In sufficiently large samples, the method will exhibit insensitivity to bias that is greater than or equal to methods, such as the Bonferroni–Holm procedure, that focus on individual outcomes; that is, Scheffé projections have larger design sensitivities. More precisely, if the least sensitive linear combination is a single one of the several outcomes, then the design sensitivity using Scheffé projections equals that using a Bonferroni correction, but if the least sensitive combination is a nontrivial combination of two or more outcomes, then Scheffé projections have larger design sensitivities. This asymptotic property is examined in terms of finite sample power of sensitivity analyses using simulation. The method is applied to a replication with recent data of a well-known study of the effects of smoking on periodontal disease. In the example, the comparison that is least sensitive to bias from unmeasured covariates combines results for lower and upper teeth, but emphasizes lower teeth. This pattern would be difficult to anticipate prior to examining the data, but Scheffé's method permits use of this unanticipated pattern without fear of capitalizing on chance.

REFERENCES

BILLINGSLEY, P. (1979). *Probability and Measure*. Wiley, New York. MR0534323

CORNFIELD, J., HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A. M., SHIMKIN, M. B. and WYNDER, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* **22** 173–203.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.

FOGARTY, C. B. and SMALL, D. S. (2016). Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *J. Amer. Statist. Assoc*. To appear, DOI:10.1080/01621459.2015.1120675.

Key words and phrases. Causal inference, design sensitivity, observational study, Scheffe projection, sensitivity analysis.

GASTWIRTH, J. L., KRIEGER, A. M. and ROSENBAUM, P. R. (2000). Asymptotic separability in sensitivity analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 545–555. MR1772414

GENZ, A. and BRETZ, F. (2009). *Computation of Multivariate Normal and t Probabilities*. *Lecture Notes in Statistics* **195**. Springer, Dordrecht. MR2840595

HANSEN, B. B. (2007). Optmatch (R package optmatch). *R News* **7** 18–24.

HELLER, R., ROSENBAUM, P. R. and SMALL, D. S. (2009). Split samples and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **104** 1090–1101. MR2750238

HODGES, J. L. JR. and LEHMANN, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *Ann. Math. Stat.* **33** 482–497. MR0156426

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6** 65–70. MR0538597

HOSMAN, C. A., HANSEN, B. B. and HOLLAND, P. W. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl. Stat.* **4** 849–870. MR2758424

HSU, J. Y. and SMALL, D. S. (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics* **69** 803–811. MR3146776

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York. MR0606374

KARLIN, S. (1992). *Mathematical Methods and Theory in Games, Programming, and Economics*. Dover Publications, New York. Vol. I: Matrix games, programming, and mathematical economics, Vol. II: The theory of infinite games, Reprint of the 1959 original. MR1160778

LEHMACHER, W., WASSMER, G. and REITMEIR, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* **47** 511–521.

LEHMANN, E. L. (1975). *Nonparametrics*. Holden Day, San Francisco.

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

LIU, W., KURAMOTO, S. J. and STUART, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prev. Sci.* **14** 570–580.

MANSKI, C. F. (1990). Nonparametric bounds on treatment effects. *Am. Econ. Rev.* **80** 319–323.

MANSKI, C. F. and NAGIN, D. S. (1990). Bounding disagreements about treatment effects: A case study of sentencing and recidivism. *Sociol. Method.* **28** 99–137.

MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056

MARITZ, J. S. (1979). A note on exact robust confidence intervals for location. *Biometrika* **66** 163–166. MR0529161

MCCANDLESS, L. C., GUSTAFSON, P. and LEVY, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Stat. Med.* **26** 2331–2347. MR2368419

MCKILLIP, J. (1992). Research without control groups: A control construct design. In *Methodological Issues in Applied Social Psychology* (F. B. Bryant et al., eds.) 159–175. Plenum Press, New York.

NEYMAN, J. (1923, 1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 [Originally published in Polish in *Ann. Agric. Sci.* **10** (1923), 1–51]. Reprinted in 1990 in *Statist. Sci.* **5** 463–464.

PITMAN, E. J. G. (1937). Significance tests that may be applied to samples from any population, I. *Supp. J. Roy. Satist. Soc.* **4** 119–130.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York. MR0346957

ROSENBAUM, P. R. (1995). Quantiles in nonrandom samples and observational studies. *J. Amer. Statist. Assoc.* **90** 1424–1431. MR1379486

ROSENBAUM, P. R. (2004). Design sensitivity in observational studies. *Biometrika* **91** 153–164. MR2050466

ROSENBAUM, P. R. (2007). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics* **63** 456–464. MR2370804

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York. MR2561612

ROSENBAUM, P. R. (2013). Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics* **69** 118–127. MR3058058

ROSENBAUM, P. R. (2014). Weighted M-statistics with superior design sensitivity in matched observational studies with multiple controls. *J. Amer. Statist. Assoc.* **109** 1145–1158. MR3265687

ROSENBAUM, P. R. (2015). Bahadur efficiency of sensitivity analyses in observational studies. *J. Amer. Statist. Assoc.* **110** 205–217. MR3338497

ROSENBAUM, P. R. and RUBIN, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Roy. Satist. Soc. B* **45** 212–218.

ROSENBAUM, P. R. and SILBER, J. H. (2009a). Amplification of sensitivity analysis in matched observational studies. *J. Amer. Statist. Assoc.* **104** 1398–1405. MR2750570

ROSENBAUM, P. R. and SILBER, J. H. (2009b). Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. *J. Amer. Statist. Assoc.* **104** 501–511. MR2751434

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Ed. Psych.* **66** 688–701.

SCHEFFÉ, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika* **40** 87–104. MR0057504

SHEPHERD, B. E., GILBERT, P. B., JEMIAI, Y. and ROTNITZKY, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics* **62** 332–342. MR2236845

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812

TOMAR, S. L. and ASMA, S. (2000). Smoking-attributable periodontitis in the United States: Findings from NHANES III. *J. Periodont.* **71** 743–751.

VAN ELTEREN, PH. (1960). On the combination of independent two sample test of Wilcoxon. *Bull. Inst. Internat. Statist.* **37** 351–361. MR0119313

WEI, L., BARKER, L. and EKE, P. (2013). Array applications in determining periodontal disease measurement. SouthEast SAS User's Group. (SESUG2013) Paper CC-15, analytics.ncsu.edu/sesug/2013/CC-15.pdf.

WEISS, N. (2002). Can the 'specificity' of an association be rehabilitated as a basis for supporting a causal hypothesis? *Epidemiology* **13** 6–8.

WELCH, B. L. (1937). On the z-test in randomized blocks and latin squares. *Biometrika* **29** 21–52.

YU, B. B. and GASTWIRTH, J. L. (2005). Sensitivity analysis for trend tests: Application to the risk of radiation exposure. *Biostatistics* **6** 201–209.

ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* **107** 1360–1371. MR3036400

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1472–1495
DOI: 10.1214/16-AOAS943

FUNCTIONAL COVARIATE-ADJUSTED PARTIAL AREA UNDER THE SPECIFICITY-ROC CURVE WITH AN APPLICATION TO METABOLIC SYNDROME DIAGNOSIS

BY VANDA INÁCIO DE CARVALHO[*,1], MIGUEL DE CARVALHO[*,1], TODD A. ALONZO[†] AND WENCESLAO GONZÁLEZ-MANTEIGA[‡,2]

Pontificia Universidad Católica de Chile[*], University of Southern California[†], and Universidade de Santiago de Compostela[‡]

Due to recent advances in technology, medical diagnosis data are becoming increasingly complex and, nowadays, applications where measurements are curves or images are ubiquitous. Motivated by the need of modeling a functional covariate on a metabolic syndrome case study, we develop a nonparametric functional regression model for the area under the specificity receiver operating characteristic curve. This partial area is a meaningful summary measure of diagnostic accuracy for cases in which misdiagnosis of diseased subjects may lead to serious clinical consequences, and hence it is critical to maintain a high sensitivity. Its normalized value can be interpreted as the average specificity over the interval of sensitivities considered, thus summarizing the trade-off between sensitivity and specificity. Our methods are motivated by, and applied to, a metabolic syndrome study that investigates how restricting the sensitivity of the gamma-glutamyl-transferase, a metabolic syndrome marker, to certain clinical meaningful values, affects its corresponding specificity and how it might change for different curves of arterial oxygen saturation. Application of our methods suggests that oxygen saturation is key to gamma-glutamyl transferase's performance and that some of the different intervals of sensitivities considered offer a good trade-off between sensitivity and specificity. The simulation study shows that the estimator associated with our model is able to recover successfully the true overall shape of the functional covariate-adjusted partial area under the curve in different complex scenarios.

REFERENCES

ADIMARI, G. and CHIOGNA, M. (2012). Jackknife empirical likelihood based confidence intervals for partial areas under ROC curves. *Statist. Sinica* **22** 1457–1477. MR3027095

ANEIROS-PÉREZ, G. and VIEU, P. (2006). Semi-functional partial linear regression. *Statist. Probab. Lett.* **76** 1102–1110. MR2269280

CAI, T. and DODD, L. E. (2008). Regression analysis for the partial area under the ROC curve. *Statist. Sinica* **18** 817–836. MR2440072

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application. Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge Univ. Press, Cambridge. MR1478673

Key words and phrases. Arterial oxygen saturation, average specificity, biomarker, functional covariate-adjustment, gamma-glutamyl transferase, kernel regression, metabolic syndrome, partial area under the curve, sensitivity, specificity-receiver operating characteristic curve.

DODD, L. E. and PEPE, M. S. (2003). Partial AUC estimation and regression. *Biometrics* **59** 614–623. MR2004266

ECKEL, R. H., GRUNDY, S. M. and ZIMMET, P. Z. (2005). The metabolic syndrome. *Lancet* **365** 1415–1428.

FEBRERO-BANDE, M. and OVIEDO DE LA FUENTE, M. (2012). Statistical computing in functional data analysis: The $\mathbb{R}$ package `fda.usc`. *J. Stat. Softw.* **51** 1–28.

FERRATY, F., VAN KEILEGOM, I. and VIEU, P. (2010). On the validity of the bootstrap in non-parametric functional regression. *Scand. J. Stat.* **37** 286–306. MR2682301

FERRATY, F. and VIEU, P. (2002). The functional nonparametric model and application to spectrometric data. *Comput. Statist.* **17** 545–564. MR1952697

FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis*: *Theory and Practice*. Springer, New York. MR2229687

GIGLIARANO, C., FIGINI, S. and MULIERE, P. (2014). Making classifier performance comparisons when ROC curves intersect. *Comput. Statist. Data Anal.* **77** 300–312. MR3210064

GONZÁLEZ-MANTEIGA, W., PARDO-FERNÁNDEZ, J. C. and VAN KEILEGOM, I. (2011). ROC curves in non-parametric location-scale regression models. *Scand. J. Stat.* **38** 169–184. MR2760145

GUDE, F., REY-GARCIA, J., FERNANDEZ-MERINO, C., MEIJIDE, L., GARCÍA-ORTIZ, L., ZAMARRON, C. and GONZALEZ-QUINTELA, A. (2009). Serum levels of gamma-glutamyl transferase are associated with markers of nocturnal hypoxemia in general adult population. *Clin. Chim. Acta* **407** 67–71.

HÄRDLE, W. (1991). *Smoothing Techniques*: *With Implementation in S*. Springer, New York. MR1140190

HÄRDLE, W. and MARRON, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19** 778–796. MR1105844

HUNG, H. and CHIANG, C.-T. (2011). Nonparametric methodology for the time-dependent partial area under the ROC curve. *J. Statist. Plann. Inference* **141** 3829–3838. MR2823653

INÁCIO, V., GONZÁLEZ-MANTEIGA, W., FEBRERO-BANDE, M., GUDE, F., ALONZO, T. A. and CADARSO-SUÁREZ, C. (2012). Extending induced ROC methodology to the functional context. *Biostat.* **13** 594–608.

INÁCIO DE CARVALHO, V., JARA, A., HANSON, T. E. and DE CARVALHO, M. (2013). Bayesian nonparametric ROC regression modeling. *Bayesian Anal.* **8** 623–645. MR3102228

INÁCIO DE CARVALHO, V., DE CARVALHO, M., ALONZO, T. A. and GONZÁLEZ-MANTEIGA, W. (2016). Supplement to "Functional covariate-adjusted partial area under the specificity-ROC curve with an application to metabolic syndrome diagnosis." DOI:10.1214/16-AOAS943SUPP.

JIANG, Y., METZ, C. E. and NISHIKAWA, R. M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* **201** 745–750.

LEE, D. S., EWANS, J. C., ROBINS, S. J., WILSON, P. W., ALBANO, I., FOX, C. S., WANG, T. J., BENJAMIN, E. J. and VASAN, R. S. (2007). Gamma glutamyl transferase and metabolic syndrome, cardiovascular disease, and mortality risk: The framingham heart study. *Arteriosclorosis, Trombosis, and Vascular Biology* **27** 127–133.

LÓPEZ-PINTADO, S. and ROMO, J. (2009). On the concept of depth for functional data. *J. Amer. Statist. Assoc.* **104** 718–734. MR2541590

MA, H., BANDOS, A. I., ROCKETTE, H. E. and GUR, D. (2013). On use of partial area under the ROC curve for evaluation of diagnostic performance. *Stat. Med.* **32** 3449–3458. MR3092242

PARDO-FERNÁNDEZ, J. C., RODRÍGUEZ-ÁLVAREZ, M. X. and VAN KEILEGOM, I. (2014). A review on ROC curves in the presence of covariates. *REVSTAT* **12** 21–41. MR3195208

$\mathbb{R}$ DEVELOPMENT CORE TEAM (2011). $\mathbb{R}$: *A Language and Environment for Statistical Computing*. $\mathbb{R}$ Foundation for Statistical Computing, Vienna.

SUN, Y. and GENTON, M. G. (2011). Functional boxplots. *J. Comput. Graph. Statist.* **20** 316–334. MR2847798

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

WANG, Z. and CHANG, Y.-C. I. (2011). Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostat*. **12** 369–385.

YAO, F., CRAIU, R. V. and REISER, B. (2010). Nonparametric covariate adjustment for receiver operating characteristic curves. *Canad. J. Statist*. **38** 27–46. MR2676928

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1496–1516
DOI: 10.1214/16-AOAS944

BAYESIAN NONPARAMETRIC DEPENDENT MODEL FOR PARTIALLY REPLICATED DATA: THE INFLUENCE OF FUEL SPILLS ON SPECIES DIVERSITY[1]

BY JULYAN ARBEL[2,*], KERRIE MENGERSEN AND JUDITH ROUSSEAU

Collegio Carlo Alberto and Bocconi University, Queensland University of Technology and Université Paris-Dauphine

We introduce a dependent Bayesian nonparametric model for the probabilistic modeling of membership of subgroups in a community based on partially replicated data. The focus here is on species-by-site data, that is, community data where observations at different sites are classified in distinct species. Our aim is to study the impact of additional covariates, for instance, environmental variables, on the data structure, and in particular on the community diversity. To this end, we introduce dependence a priori across the covariates and show that it improves posterior inference. We use a dependent version of the Griffiths–Engen–McCloskey distribution defined via the stick-breaking construction. This distribution is obtained by transforming a Gaussian process whose covariance function controls the desired dependence. The resulting posterior distribution is sampled by Markov chain Monte Carlo. We illustrate the application of our model to a soil microbial data set acquired across a hydrocarbon contamination gradient at the site of a fuel spill in Antarctica. This method allows for inference on a number of quantities of interest in ecotoxicology, such as diversity or effective concentrations, and is broadly applicable to the general problem of community response to environmental variables.

REFERENCES

AITCHISON, J. (1982). The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B* **44** 139–177. MR0676206

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London. MR0865647

AITCHISON, J. (1994). Principles of compositional data analysis. In *Multivariate Analysis and Its Applications* (*Hong Kong*, 1992). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **24** 73–81. IMS, Hayward, CA. MR1479457

ALSTON, C. L., MENGERSEN, K. L. and GARDNER, G. E. (2011). Bayesian mixture models: A blood-free dissection of a sheep. In *Mixtures*: *Estimation and Applications* (K. Mengersen, C. P. Robert and M. Titterington, eds.) 293–308. Wiley, Chichester. MR2883358

ANDRIANAKIS, I. and CHALLENOR, P. G. (2012). The effect of the nugget on Gaussian process emulators of computer models. *Comput. Statist. Data Anal.* **56** 4215–4228. MR2957866

ARBEL, J. (2013). Contributions to Bayesian nonparametric statistics. Ph.D. thesis, Univ. Paris-Dauphine.

Key words and phrases. Bayesian nonparametrics, covariate-dependent model, Gaussian processes, Griffiths–Engen–McCloskey distribution, partially replicated data, stick-breaking representation.

ARBEL, J., MENGERSEN, K. and ROUSSEAU, J. (2016). Supplement to "Bayesian nonparametric dependent model for partially replicated data: The influence of fuel spills on species diversity." DOI:10.1214/16-AOAS944SUPP.

ARBEL, J., KING, C. K., RAYMOND, B., WINSLEY, T. and MENGERSEN, K. L. (2015). Application of a Bayesian nonparametric model to derive toxicity estimates based on the response of Antarctic microbial communities to fuel-contaminated soil. *Ecol. Evol.* **5** 2633–2645.

ARBEL, J., FAVARO, S., NIPOTI, B. and TEH, Y. W. (2016). Bayesian nonparametric inference for discovery probabilities: Credible intervals and large sample asymptotics. *Statist. Sinica*. To appear. Available at arXiv:1506.04915.

BARRIENTOS, A. F., JARA, A. and QUINTANA, F. A. (2012). On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Anal.* **7** 277–309. MR2934952

BARRIENTOS, A. F., JARA, A. and QUINTANA, F. A. (2015). Bayesian density estimation for compositional data using random Bernstein polynomials. *J. Statist. Plann. Inference* **166** 116–125. MR3390138

BOHLIN, J., SKJERVE, E. and USSERY, D. (2009). Analysis of genomic signatures in prokaryotes using multinomial regression and hierarchical clustering. *BMC Genomics* **10** 487.

BORGES, E. P. and RODITI, I. (1998). A family of nonextensive entropies. *Phys. Lett. A* **246** 399–402. MR1649464

BROMS, K. M., HOOTEN, M. B. and FITZPATRICK, R. M. (2015). Accounting for imperfect detection in Hill numbers for biodiversity studies. *Methods in Ecology and Evolution* **6** 99–108.

CALABRESE, E. J. (2005). Paradigm lost, paradigm found: The re-emergence of hormesis as a fundamental dose response model in the toxicological sciences. *Environ. Pollut.* **138** 378–411.

CARON, F., DAVY, M. and DOUCET, A. (2007). Generalized Pólya urn for time-varying Dirichlet process mixtures. In *23rd Conference on Uncertainty in Artificial Intelligence* (*UAI'*2007). Vancouver, Canada.

CERQUETTI, A. (2014). Bayesian nonparametric estimation of Patil–Taillie–Tsallis diversity under Gnedin–Pitman priors. Preprint. Available at arXiv:1404.3441.

CHUNG, Y. and DUNSON, D. B. (2011). The local Dirichlet process. *Ann. Inst. Statist. Math.* **63** 59–80. MR2748934

COLWELL, R. K., CHAO, A., GOTELLI, N. J., LIN, S.-Y., MAO, C. X., CHAZDON, R. L. and LONGINO, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology* **5** 3–21.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York. MR1239641

DE'ATH, G. (2012). The multinomial diversity model: Linking Shannon diversity to multiple predictors. *Ecology* **93** 2286–2296.

DONNELLY, P. and GRIMMETT, G. (1993). On the asymptotic distribution of large prime factors. *J. Lond. Math. Soc.* (2) **47** 395–404. MR1214904

DORAZIO, R. M., MUKHERJEE, B., ZHANG, L., GHOSH, M., JELKS, H. L. and JORDAN, F. (2008). Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics* **64** 635–644, 670–671. MR2432438

DUNSON, D. B. and PARK, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* **95** 307–323. MR2521586

DUNSON, D. B., PILLAI, N. and PARK, J.-H. (2007). Bayesian density regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 163–183. MR2325270

DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Amer. Statist. Assoc.* **104** 1042–1051. MR2562004

DUNSTAN, P. K., FOSTER, S. D. and DARNELL, R. (2011). Model based grouping of species across environmental gradients. *Ecol. Model.* **222** 955–963.

ELLIS, N., SMITH, S. J. and PITCHER, C. R. (2011). Gradient forests: Calculating importance gradients on physical predictors. *Ecology* **93** 156–168.

FAVARO, S., LIJOI, A. and PRÜNSTER, I. (2012). A new estimator of the discovery probability. *Biometrics* **68** 1188–1196. MR3040025

FAVARO, S., NIPOTI, B. and TEH, Y. W. (2016). Rediscovery of Good–Turing estimators via Bayesian nonparametrics. *Biometrics* **72** 136–145.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

FERRIER, S. and GUISAN, A. (2006). Spatial modelling of biodiversity at the community level. *J. Appl. Ecol.* **43** 393–404.

FERRIER, S., MANION, G., ELITH, J. and RICHARDSON, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Divers. Distrib.* **13** 252–264.

FORDYCE, J. A., GOMPERT, Z., FORISTER, M. L. and NICE, C. C. (2011). A hierarchical Bayesian approach to ecological count data: A flexible tool for ecologists. *PLoS ONE* **6** e26785.

FOSTER, S. D. and DUNSTAN, P. K. (2010). The analysis of biodiversity using rank abundance distributions. *Biometrics* **66** 186–195. MR2756705

GELFAND, A. E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* 145–161. Chapman & Hall, London. MR1397969

GEORGE, A. W., MENGERSEN, K. and DAVIS, G. P. (2000). Localization of a quantitative trait locus via a Bayesian approach. *Biometrics* **56** 40–51.

GIBBS, M. N. (1997). Bayesian Gaussian processes for regression and classification. Ph.D. thesis, Citeseer.

GILL, C. A. and JOANES, D. N. (1979). Bayesian estimation of Shannon's index of diversity. *Biometrika* **66** 81–85. MR0529150

GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264. MR0061330

GRIFFIN, J. E. and STEEL, M. F. J. (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 179–194. MR2268037

GRIFFIN, J. E. and STEEL, M. F. J. (2011). Stick-breaking autoregressive processes. *J. Econometrics* **162** 383–396. MR2795625

HAVRDA, J. and CHARVÁT, F. (1967). Quantification method of classification processes. Concept of structural a-entropy. *Kybernetika* (*Prague*) **3** 30–35. MR0209067

HILL, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology* **54** 427–432.

HOLMES, I., HARRIS, K. and QUINCE, C. (2012). Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS ONE* **7** e30126.

JOHNSON, D. S., REAM, R. R., TOWELL, R. G., WILLIAMS, M. T. and LEON GUERRERO, J. D. (2013). Bayesian clustering of animal abundance trends for inference and dimension reduction. *J. Agric. Biol. Environ. Stat.* **18** 299–313. MR3110895

KANIADAKIS, G., LISSIA, M. and SCARFONE, A. M. (2005). Two-parameter deformations of logarithm, exponential, and entropy: A consistent framework for generalized statistical mechanics. *Phys. Rev. E* (3) **71** 046128, 12. MR2139991

LI, H. (2015). Microbiome, metagenomics and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application* **2** 73–94.

LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94** 769–786. MR2416792

LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014a). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20** 1260–1291. MR3217444

LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014b). Dependent mixture models: Clustering and borrowing information. *Comput. Statist. Data Anal.* **71** 417–433. MR3131980

LOVELL, D., PAWLOWSKY-GLAHN, V., EGOZCUE, J. J., MARGUERAT, S. and BÄHLER, J. (2015). Proportionality: A valid alternative to correlation for relative data. *PLoS Comput. Biol.* **11** e1004075.

MACEACHERN, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science* 50–55. Amer. Statist. Assoc., Alexandria, VA.

MACEACHERN, S. N. (2000). Dependent Dirichlet processes. Technical report, Dept. Statistics, The Ohio State Univ.

NEWMAN, M. C. (2012). *Quantitative Ecotoxicology*. CRC Press, Boca Raton, FL.

PATI, D., DUNSON, D. B. and TOKDAR, S. T. (2013). Posterior consistency in conditional distribution estimation. *J. Multivariate Anal.* **116** 456–472. MR3049916

PATIL, G. P. and TAILLIE, C. (1982). Diversity as a concept and its measurement. *J. Amer. Statist. Assoc.* **77** 548–567. MR0675883

PAWLOWSKY-GLAHN, V. and BUCCIANTI, A., eds. (2011). *Compositional Data Analysis*: *Theory and Applications*. Wiley, Chichester. MR2920574

PITMAN, J. (2006). *Combinatorial Stochastic Processes*. *Lecture Notes in Math.* **1875**. Springer, Berlin. MR2245368

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

RODRÍGUEZ, A. and DUNSON, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal.* **6** 145–177. MR2781811

RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2010). Latent stick-breaking processes. *J. Amer. Statist. Assoc.* **105** 647–659. MR2724849

ROYLE, J. A. and DORAZIO, R. M. (2006). Hierarchical models of animal abundance and occurrence. *J. Agric. Biol. Environ. Stat.* **11** 249–263.

ROYLE, J. A. and DORAZIO, R. M. (2008). *Hierarchical Modeling and Inference in Ecology*: *The Analysis of Data from Populations*, *Metapopulations and Communities*. Academic Press, San Diego, CA.

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

SCHLOSS, P. D., WESTCOTT, S. L., RYABIN, T., HALL, J. R., HARTMANN, M., HOLLISTER, E. B., LESNIEWSKI, R. A., OAKLEY, B. B., PARKS, D. H., ROBINSON, C. J., SAHL, J. W., STRES, B., THALLINGER, G. G., HORN, D. J. V. and WEBER, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75** 7537–7541.

SICILIANO, S. D., PALMER, A. S., WINSLEY, T., LAMB, E., BISSETT, A., BROWN, M. V., VAN DORST, J., JI, M., FERRARI, B. C., GROGAN, P., CHU, H. and SNAPE, I. (2014). Soil fertility is associated with fungal and bacterial richness, whereas pH is associated with community composition in polar soil microbial communities. *Soil Biol. Biochem.* **78** 10–20.

SNAPE, I., SICILIANO, S. D., WINSLEY, T., VAN DORST, J., MUKAN, J., PALMER, A. S. and LAGEREWSKIJ, G. (2015). *Operational Taxonomic Unit* (*OTU*) *Microbial Ecotoxicology Data from Macquarie Island and Casey Station*: *TPH*, *Chemistry and OTU Abundance Data*. Australian Antarctic Data Centre.

VAN DEN BOOGAART, K. G. and TOLOSANA-DELGADO, R. (2013). Fundamental concepts of compositional data analysis. In *Analyzing Compositional Data with R*, *Use R!*. Springer, Heidelberg. MR3099409

VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442

WANG, Y., NAUMANN, U., WRIGHT, S. T. and WARTON, D. I. (2012). mvabund—An R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* **3** 471–474.

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1517–1546
DOI: 10.1214/16-AOAS945

BAYESIAN DATA FUSION APPROACHES TO PREDICTING SPATIAL TRACKS: APPLICATION TO MARINE MAMMALS

BY YANG LIU[1], JAMES V. ZIDEK[1], ANDREW W. TRITES AND BRIAN C. BATTAILE

University of British Columbia

Bayesian Melding (BM) and downscaling are two Bayesian approaches commonly used to combine data from different sources for statistical inference. We extend these two approaches to combine accurate but sparse direct observations with another set of high-resolution but biased calculated observations. We use our methods to estimate the path of a moving or evolving object and apply them in a case study of tracking northern fur seals. To make the BM approach computationally feasible for high-dimensional (big) data, we exploit the properties of the processes along with approximations to the likelihood to break the high-dimensional problem into a series of lower dimensional problems. To implement the alternative, downscaling approach, we use R-INLA to connect the two sources of observations via a linear mixed effect model. We compare the predictions of the two approaches by cross-validation as well as simulations. Our results show that both approaches yield similar results—both provide accurate, high resolution estimates of the at-sea locations of the northern fur seals, as well as Bayesian credible intervals to characterize the uncertainty about the estimated movement paths.

REFERENCES

BATTAILE, B. (2014). TrackReconstruction: Reconstruct animal tracks from magnetometer, accelerometer, depth and optional speed data. R package version 1.1.

BATTAILE, B. C., NORDSTROM, C. A., LIEBSCH, N. and TRITES, A. W. (2015). Foraging a new trail with northern fur seals (Callorhinus ursinus): Lactating seals from islands with contrasting population dynamics have different foraging strategies, and forage at scales previously unrecognized by GPS interpolated dive data. *Marine Mammal Science* **31** 1494–1520.

BENOIT-BIRD, K. J., BATTAILE, B. C., HEPPELL, S. A., HOOVER, B., IRONS, D., JONES, N., KULETZ, K. J., NORDSTROM, C. A., PAREDES, R., SURYAN, R. M., WALUK, C. M. and TRITES, A. W. (2013a). Prey patch patterns predict habitat use by top marine predators with diverse foraging strategies. *PLoS ONE* **8** e53348.

BENOIT-BIRD, K. J., BATTAILE, B. C., NORDSTROM, C. A. and TRITES, A. W. (2013b). Foraging behavior of northern fur seals closely matches the hierarchical patch scales of prey. *Mar. Ecol. Prog. Ser.* **479** 283–302.

BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *J. Agric. Biol. Environ. Stat.* **15** 176–197. MR2787270

BLOCK, B. A., JONSEN, I. D., JORGENSEN, S. J., WINSHIP, A. J., SHAFFER, S. A., BOGRAD, S. J., HAZEN, E. L., FOLEY, D. G., BREED, G. A., HARRISON, A.-L., GANONG, J. E.,

Key words and phrases. Bayesian melding, downscaling, bio-logging, conditional independence, INLA, Dead-Reckoning, tracking, marine mammals, Northern fur seal.

SWITHENBANK, A., CASTLETON, M., DEWAR, H., MATE, B. R., SHILLINGER, G. L., SCHAEFER, K. M., BENSON, S. R., WEISE, M. J., HENRY, R. W. and COSTA, D. P. (2011). Tracking apex marine predator movements in a dynamic ocean. *Nature* **475** 86–90.

BOEHME, L., KOVACS, K. and LYDERSEN, C. (2010). Biologging in the global ocean observing system. *Proceedings of OceanObs* 09: *Sustained Ocean Observations and Information for Society* **2** 21–25.

BOEHME, L., MEREDITH, M. P., THORPE, S. E., BIUW, M. and FEDAK, M. (2008). Antarctic circumpolar current frontal system in the South Atlantic: Monitoring using merged Argo and animal-borne sensor data. *Journal of Geophysical Research C*: *Oceans* **113** C9.

BRYANT, E. (2007). 2D Location Accuracy Statistics for Fastloc RCores Running Firmware Versions 2.2 and 2.3. Wildtrack Telemetry Systems Ltd.

DUGAS, A. F., JALALPOUR, M., GEL, Y., LEVIN, S., TORCASO, F., IGUSA, T. and ROTHMAN, R. E. (2013). Influenza forecasting with Google Flu Trends. *PLoS ONE* **8** e56176.

DUKIC, V., LOPES, H. F. and POLSON, N. G. (2012). Tracking epidemics with Google Flu Trends data and a state-space SEIR model. *J. Amer. Statist. Assoc.* **107** 1410–1426. MR3036404

ELKAIM, G. H., DECKER, E. B., OLIVER, G. and WRIGHT, B. (2006). Marine Mammal Marker (MAMMARK) dead reckoning sensor for in-situ environmental monitoring. *Proceedings of the ION/IEEE PLANS* 976–987.

FLEMING, C. H., FAGAN, W. F., MUELLER, T., OLSON, K. A., LEIMGRUBER, P. and CALABRESE, J. M. (2016). Estimating where and how animals travel: An optimal framework for path reconstruction from autocorrelated tracking data. *Ecology*.

FOLEY, K. M. and FUENTES, M. (2008). A statistical framework to combine multivariate spatial data and physical models for hurricane surface wind prediction. *J. Agric. Biol. Environ. Stat.* **13** 37–59. MR2423075

FUENTES, M. and RAFTERY, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61** 36–45. MR2129199

GINSBERG, J., MOHEBBI, M. H., PATEL, R. S., BRAMMER, L., SMOLINSKI, M. S. and BRILLIANT, L. (2009). Detecting influenza epidemics using search engine query data. *Nature* **457** 1012–1014.

HENDERSON, H. V. and SEARLE, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Rev.* **23** 53–60. MR0605440

HOOTEN, M. B., HANKS, E. M., JOHNSON, D. S. and ALLDREDGE, M. W. (2013). Reconciling resource utilization and resource selection functions. *J. Anim. Ecol.* **82** 1146–1154.

HORNE, J. S., GARTON, E. O., KRONE, S. M. and LEWIS, J. S. (2007). Analyzing animal movements using Brownian bridges. *Ecology* **88** 2354–2363.

HUMPHRIES, N. E., QUEIROZ, N., DYER, J. R. M., PADE, N. G., MUSYL, M. K., SCHAEFER, K. M., FULLER, D. W., BRUNNSCHWEILER, J. M., DOYLE, T. K., HOUGHTON, J. D. R., HAYS, G. C., JONES, C. S., NOBLE, L. R., WEARMOUTH, V. J., SOUTHALL, E. J. and SIMS, D. W. (2010). Environmental context explains Lévy and Brownian movement patterns of marine predators. *Nature* **465** 1066–1069.

JOHNSON, M. P. and TYACK, P. L. (2003). A digital acoustic recording tag for measuring the response of wild marine mammals to sound. *IEEE Journal of Oceanic Engineering* **28** 3–12.

JOHNSON, D. S., LONDON, J. M., LEA, M.-A. and DURBAN, J. W. (2008). Continuous-time correlated random walk model for animal telemetry data. *Ecology* **89** 1208–1215.

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

KRANSTAUBER, B., SAFI, K. and BARTUMEUS, F. (2014). Bivariate Gaussian bridges: Directional factorization of diffusion in Brownian bridge models. *Mov. Ecol.* **2** 5.

KRANSTAUBER, B., KAYS, R., LAPOINT, S. D., WIKELSKI, M. and SAFI, K. (2012). A dynamic Brownian bridge movement model to estimate utilization distributions for heterogeneous animal movement. *J. Anim. Ecol.* **81** 738–746.

LAZER, D., KENNEDY, R., KING, G. and VESPIGNANI, A. (2014a). Big data. The parable of Google Flu: Traps in big data analysis. *Science* **343** 1203–1205.

LAZER, D., KENNEDY, R., KING, G. and VESPIGNANI, A. (2014b). Google Flu Trends still appears sick: An evaluation of the 2013-2014 Flu season. *SSRN Electronic Journal* 1–11.

LE, N. D. and ZIDEK, J. V. (2006). *Statistical Analysis of Environmental Space–Time Processes*. Springer, New York. MR2223933

LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. MR2853727

LIU, Y. (2014). BayesianAnimalTracker: Bayesian Melding of GPS and DR Path for Animal Tracking. R package version 1.2.

LIU, Z., LE, N. D. and ZIDEK, J. V. (2011). An empirical assessment of Bayesian melding for mapping ozone pollution. *Environmetrics* **22** 340–353. MR2843389

LIU, Y., BATTAILE, B. C., ZIDEK, J. V. and TRITES, A. W. (2015). Bias correction and uncertainty characterization of dead-reckoned paths of marine mammals. *Animal Biotelmetry* **3** 51.

LIU, Y., DINSDALE, D. R., JUN, S.-H., BRIERCLIFFE, C. and BONE, J. (2016a). Automatic learning of basketball strategy via SportVU tracking data: The potential field approach. Cascadia Symposium on Statistics in Sports, Vancouver.

LIU, Y., ZIDEK, J. V., TRITES, A. W. and BATTAILE, B. C., (2016b). Supplement to "Bayesian data fusion approaches to predicting spatial tracks: Application to marine mammals." DOI:10.1214/16-AOAS945SUPP.

MARTINS, T. G., SIMPSON, D., LINDGREN, F. and RUE, H. (2013). Bayesian computing with INLA: New features. *Comput. Statist. Data Anal.* **67** 68–83. MR3079584

MCCLINTOCK, B. T., JOHNSON, D. S., HOOTEN, M. B., VER HOEF, J. M. and MORALES, J. M. (2014). When to be discrete: The importance of time formulation in understanding animal movement. *Mov. Ecol.* **2** 1–14.

MILLER, A., BORNN, L., ADAMS, R. and GOLDSBERRY, K. (2014). Factorized point process intensities: A spatial analysis of professional basketball. Preprint. Available at arXiv:1401.0942.

MITANI, Y., SATO, K., ITO, S., CAMERON, M. F., SINIFF, D. B. and NAITO, Y. (2003). A method for reconstructing three-dimensional dive profiles of marine mammals using geomagnetic intensity data: Results from two lactating Weddell seals. *Polar Biology* **26** 311–317.

NORDSTROM, C. A., BENOIT-BIRD, K. J., BATTAILE, B. C. and TRITES, A. W. (2013). Northern fur seals augment ship-derived ocean temperatures with higher temporal and spatial resolution data in the eastern Bering Sea. *Deep Sea Research Part II* **94** 257–273.

POZDNYAKOV, V., MEYER, T., WANG, Y.-B. and YAN, J. (2014). On modeling animal movements using Brownian motion with measurement error. *Ecology* **95** 247–253.

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

RUNDEL, C. W., SCHLIEP, E. M., GELFAND, A. E. and HOLLAND, D. M. (2015). A data fusion approach for spatial analysis of speciated PM2.5 across time. *Environmetrics*.

SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. MR1041765

SAHU, S. K., GELFAND, A. E. and HOLLAND, D. M. (2010). Fusing point and areal level space-time data with application to wet deposition. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59** 77–103. MR2750133

SALZBERG, S. (2014). Why Google flu is a failure. Forbes.com [online] 03–24.

SAWYER, H., KAUFFMAN, M. J., NIELSON, R. M. and HORNE, J. S. (2009). Identifying and prioritizing ungulate migration routes for landscape-level conservation. *Ecological Applications* **19** 2016–2025.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. MR1979380

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2014). The deviance information criterion: 12 years on. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 485–493. MR3210727

STIRZAKER, G. G. D. and GRIMMETT, D. (2001). *Probability and Random Processes*. Oxford Science Publications, Oxford.

WAHBA, G. (1965). A least squares estimate of satellite attitude. *SIAM Rev.* **7** 409–409.

WILSON, R. P. and WILSON, M. P. (1988). Dead reckoning–a new technique for determining penguin movements at sea. *Meeresforschung–Reports on Marine Research* **32** 155–158.

WILSON, R. P., LIEBSCH, N., DAVIES, I. M., QUINTANA, F., WEIMERSKIRCH, H., STORCH, S., LUCKE, K., SIEBERT, U., ZANKL, S., MÜLLER, G., ZIMMER, I., SCOLARO, A., CAMPAGNA, C., PLÖTZ, J., BORNEMANN, H., TEILMANN, J. and MCMAHON, C. R. (2007). All at sea with animal tracks; methodological and analytical solutions for the resolution of movement. *Deep Sea Research Part II* **54** 193–210.

ZIDEK, J. V., LE, N. D. and LIU, Z. (2012). Combining data and simulated data for space-time fields: Application to ozone. *Environ. Ecol. Stat.* **19** 37–56. MR2909084

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1547–1571
DOI: 10.1214/16-AOAS948

A BAYESIAN PREDICTIVE MODEL FOR IMAGING GENETICS WITH APPLICATION TO SCHIZOPHRENIA[1]

BY THIERRY CHEKOUO[*], FRANCESCO C. STINGO[†,1], MICHELE GUINDANI[‡,1] AND KIM-ANH DO[§,1]

University of Minnesota Duluth[*], University of Florence[†], University of California, Irvine[‡], and University of Texas MD Anderson Cancer Center[§]

Imaging genetics has rapidly emerged as a promising approach for investigating the genetic determinants of brain mechanisms that underlie an individual's behavior or psychiatric condition. In particular, for early detection and targeted treatment of schizophrenia, it is of high clinical relevance to identify genetic variants and imaging-based biomarkers that can be used as diagnostic markers, in addition to commonly used symptom-based assessments. By combining single-nucleotide polymorphism (SNP) arrays and functional magnetic resonance imaging (fMRI), we propose an integrative Bayesian risk prediction model that allows us to discriminate between individuals with schizophrenia and healthy controls, based on a sparse set of discriminatory regions of interest (ROIs) and SNPs. Inference on a regulatory network between SNPs and ROI intensities (ROI–SNP network) is used in a single modeling framework to inform the selection of the discriminatory ROIs and SNPs. We use simulation studies to assess the performance of our method and apply it to data collected from individuals with schizophrenia and healthy controls. We found our approach to outperform competing methods that do not link the ROI–SNP network to the selection of discriminatory markers.

REFERENCES

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

ATCHADÉ, Y. F., LARTILLOT, N. and ROBERT, C. (2013). Bayesian computation for statistical models with intractable normalizing constants. *Braz. J. Probab. Stat.* **27** 416–436. MR3105037

BATMANGHELICH, N., DALCA, A., SABUNCU, M. and GOLLAND, P. (2013). Joint modeling of imaging and genetics. In *Information Processing in Medical Imaging* (J. C. Gee, S. Joshi, K. Pohl, W. M. Wells and L. Zellei, eds.). *Lecture Notes in Computer Science* **7917** 766–777. Springer, Berlin.

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

BOWMAN, F. D. (2014). Brain imaging analysis. *Annu. Rev. Stat. Appl.* **1** 61–85.

CALHOUN, V. D. and HUGDAHL, K. (2012). Cognition and neuroimaging in schizophrenia. *Front. Human Neurosci.* **6** 276.

CALHOUN, V. D., LIU, J. and ADALI, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage* **45** S163–S172.

Key words and phrases. Imaging genetics, fMRI, data integration, Bayesian variable selection, Markov random field, nonlocal prior.

CANNON, T. D. and KELLER, M. C. (2006). Endophenotypes in the genetic analyses of mental disorders. *Annu. Rev. Clin. Psychol.* **2** 267–290.

CAO, H., DUAN, J., LIN, D., CALHOUN, V. and WANG, Y.-P. (2013). Integrating fMRI and SNP data for biomarker identification for schizophrenia with a sparse representation based variable selection method. *BMC Medical Genomics* **6 Suppl 3** S2.

CHEKOUO, T., STINGO, F. C., GUINDANI, M. and DO, K. (2016). Supplement to "A Bayesian predictive model for imaging genetics with application to schizophrenia." DOI:10.1214/16-AOAS948SUPP.

CHEN, J., CALHOUN, V. D., PEARLSON, G. D., EHRLICH, S., TURNER, J. A., HO, B.-C., WASSINK, T. H., MICHAEL, A. and LIU, J. (2012). Multifaceted genomic risk for brain function in schizophrenia. *NeuroImage* **61** 866–875.

CHI, E. C., ALLEN, G. I., ZHOU, H., KOHANNIM, O., LANGE, K. and THOMPSON, P. M. (2013). Imaging genetics via sparse canonical correlation analysis. In *Biomedical Imaging* (*ISBI*), 2013 *IEEE* 10*th International Symposium on* 740–743.

DAMARAJU, E., ALLEN, E. A., BELGER, A., FORD, J. M., MCEWEN, S., MATHALON, D. H., CALHOUN, V. D. et al. (2014). Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage*: *Clinical* **5** 298–308.

DENG, L. and YU, D. (2013). Deep learning: Methods and applications. *Found. Trends Signal Process.* **7** 197–391. MR3295556

DETTLING, M. and BÜHLMANN, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19** 1061–1069.

ERHARDT, E. B., RACHAKONDA, S., BEDRICK, E. J., ALLEN, E. A., ADALI, T. and CALHOUN, V. D. (2011). Comparison of multi-subject ICA methods for analysis of fMRI data. *Hum. Brain Mapp.* **32** 2075–2095.

FILIPOVYCH, R., RESNICK, S. M. and DAVATZIKOS, C. (2011). Multi-kernel classification for integration of clinical and imaging data: Application to prediction of cognitive decline in older adults machine learning in medical imaging (K. Suzuki, F. Wang, D. Shen and P. Yan, eds.). *Lecture Notes in Computer Science* **7009** 26–34. Springer, Berlin.

FLOCH, É. L., GUILLEMOT, V., FROUIN, V., PINEL, P., LALANNE, C., TRINCHERA, L., TENENHAUS, A., MORENO, A., ZILBOVICIUS, M., BOURGERON, T., DEHAENE, S., THIRION, B., POLINE, J.-B. and DUCHESNAY, É. (2012). Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *NeuroImage* **63** 11–24.

FORNITO, A., ZALESKY, A., PANTELIS, C. and BULLMORE, E. T. (2012). Schizophrenia, neuroimaging and connectomics. *NeuroImage* **62** 2296–2314.

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

FRIEDMAN, J. H., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

FUJIWARA, H., NAMIKI, C., HIRAO, K., MIYATA, J., SHIMIZU, M., FUKUYAMA, H., SAWAMOTO, N., HAYASHI, T. and MURAI, T. (2007). Anterior and posterior cingulum abnormalities and their association with psychopathology in schizophrenia: A diffusion tensor imaging study. *Schizophr. Res.* **95** 215–222.

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–374.

GLAHN, D. C., LAIRD, A. R., ELLISON-WRIGHT, I., THELEN, S. M., ROBINSON, J. L., LANCASTER, J. L., BULLMORE, E. and FOX, P. T. (2008). Meta-analysis of gray matter anomalies in schizophrenia: Application of anatomic likelihood estimation and network analysis. *Biological Psychiatry* **64** 774–781.

GOLDSMITH, J., HUANG, L. and CRAINICEANU, C. M. (2014). Smooth scalar-on-image regression via spatial Bayesian variable selection. *J. Comput. Graph. Statist.* **23** 46–64. MR3173760

GOLDSMITH, J., CRAINICEANU, C. M., CAFFO, B. and REICH, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 453–469. MR2914521

GOLLUB, R. L., SHOEMAKER, J. M., KING, M. D., WHITE, T., EHRLICH, S., SPONHEIM, S. R., CLARK, V. P., TURNER, J. A., MUELLER, B. A., MAGNOTTA, V. et al. (2013). The MCIC collection: A shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics* **11** 367–388.

HARDOON, D. R., ETTINGER, U., MOURÃO-MIRANDA, J., ANTONOVA, E., COLLIER, D., KUMARI, V., WILLIAMS, S. C. R. and BRAMMER, M. (2009). Correlation-based multivariate analysis of genetic influence on brain volume. *Neurosci. Lett.* **450** 281–286.

HARIRI, A. R. and WEINBERGER, D. R. (2003). Imaging genomics. *Br. Med. Bull.* **65** 259–270.

IKEDA, M., YAMANOUCHI, Y., KINOSHITA, Y., KITAJIMA, T., YOSHIMURA, R., HASHIMOTO, S., O'DONOVAN, M. C., NAKAMURA, J., OZAKI, N. and IWATA, N. (2008). Variants of dopamine and serotonin candidate genes as predictors of response to risperidone treatment in first-episode schizophrenia. *Pharmacogenomics* **9** 1437–1443.

JACOB, A. (2013). Limitations of clinical psychiatric diagnostic measurements. *J. Neurol. Disord.* **2**.

JOHNSON, V. E. and ROSSELL, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 143–170. MR2830762

JOHNSON, V. E. and ROSSELL, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. MR2980074

JOO, E. J., LEE, K. Y., JEONG, S. H., ROH, M. S., KIM, S. H., AHN, Y. M. and KIM, Y. S. (2009). AKT1 gene polymorphisms and obstetric complications in the patients with schizophrenia. *Psychiatry Investigation* **6** 102–107.

KIM, J. Y., LIU, C. Y., ZHANG, F., DUAN, X., WEN, Z., SONG, J., FEIGHERY, E., LU, B., RUJESCU, D., CLAIR, D. S., CHRISTIAN, K., CALLICOTT, J. H., WEINBERGER, D. R., SONG, H. and LI MING, G. (2012). Interplay between DISC1 and GABA signaling regulates neurogenesis in mice and risk for schizophrenia. *Cell* **148** 1051–1064.

LECUN, Y., BENGIO, Y. and HINTON, G. (2015). Deep learning. *Nature* **521** 436–444.

LENCZ, T., MORGAN, T. V., ATHANASIOU, M., DAIN, B., REED, C. R., KANE, J. M., KUCHERLAPATI, R. and MALHOTRA, A. K. (2007). Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. *Mol. Psychiatry* **12** 572–580.

LEVITT, J. J., MCCARLEY, R. W., NESTOR, P. G., PETRESCU, C., DONNINO, R., HIRAYASU, Y., KIKINIS, R., JOLESZ, F. A. and SHENTON, M. E. (1999). Quantitative volumetric MRI study of the cerebellum and vermis in schizophrenia: Clinical and cognitive correlates. *Am. J. Psychiatr.* **156** 1105–1107.

LI, F. and ZHANG, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *J. Amer. Statist. Assoc.* **105** 1202–1214. MR2752615

LI, F., ZHANG, T., WANG, Q., GONZALEZ, M. Z., MARESH, E. L. and COAN, J. (2015). Spatial Bayesian variable selection and grouping in high-dimensional scalar-on-image regressions. *Ann. Appl. Stat.* **9** 687–713.

LIN, D., CALHOUN, V. D. and WANG, Y.-P. (2014). Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. *Med. Image Anal.* **18** 891–902.

LIN, J.-A., ZHU, H., MIHYE, A., SUN, W., IBRAHIM, J. G. and FOR THE ALZHEIMER'S NEUROIMAGING INITIATIVE (2014). Functional-mixed effects models for candidate genetic mapping in imaging genetic studies. *Genet. Epidemiol.* **38** 680–691.

LINDQUIST, M. A. (2008). The statistical analysis of fMRI data. *Statist. Sci.* **23** 439–464. MR2530545

LIU, J. and CALHOUN, V. D. (2014). A review of multivariate analyses in imaging genetics. *Front. Neuroinform.* **8** 29.

LIU, J., PEARLSON, G., WINDEMUTH, A., RUANO, G., PERRONE-BIZZOZERO, N. I. and CALHOUN, V. (2009). Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Hum. Brain Mapp.* **30** 241–255.

LO, W.-S., LAU, C.-F., XUAN, Z., CHAN, C.-F., FENG, G.-Y., HE, L., CAO, Z.-C., LIU, H., LUAN, Q.-M. and XUE, H. (2004). Association of SNPs and haplotypes in GABAA receptor beta2 gene with schizophrenia. *Mol. Psychiatry* **9** 603–608.

MADIGAN, D. and RAFTERY, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89** 1535–1546.

MEDA, S. A., NARAYANAN, B., LIU, J., PERRONE-BIZZOZERO, N. I., STEVENS, M. C., CALHOUN, V. D., GLAHN, D. C., SHEN, L., RISACHER, S. L., SAYKIN, A. J. and PEARLSON, G. D. (2012). A large scale multivariate parallel {ICA} method reveals novel imaging–genetic relationships for Alzheimer's disease in the {ADNI} cohort. *NeuroImage* **60** 1608–1621.

MEYER-LINDENBERG, A. (2012). The future of fMRI and genetics research. *NeuroImage* **62** 1286–1292.

MÜLLER, V. I., CIESLIK, E. C., LAIRD, A. R., FOX, P. T. and EICKHOFF, S. B. (2013). Dysregulated left inferior parietal activity in schizophrenia and depression: Functional connectivity and characterization. *Front. Human Neurosci.* **7** 268.

OKUGAWA, G., SEDVALL, G. C. and AGARTZ, I. (2003). Smaller cerebellar vermis but not hemisphere volumes in patients with chronic schizophrenia. *Am. J. Psychiatr.* **160** 1614–1617.

POTKIN, S. G., TURNER, J. A., FALLON, J. A., LAKATOS, A., KEATOR, D. B., GUFFANTI, G. and MACCIARDI, F. (2009). Gene discovery through imaging genetics: Identification of two novel genes associated with schizophrenia. *Mol. Psychiatry* **14** 416–428.

POTKIN, S. G., VAN ERP, T. G. M., LING, S., MACCIARDI, F. and XIE, X. (2015). Identifying Unanticipated Genes and Mechanisms in Serious Mental Illness: GWAS Based Imaging Genetics Strategies. 209. Oxford Univ. Press, London.

RIPLEY, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge Univ. Press, Cambridge. MR1438788

SAETRE, P., AGARTZ, I., FRANCISCIS, A. D., LUNDMARK, P., DJUROVIC, S., KAHLER, A., ANDREASSEN, O. A., JAKOBSEN, K. D., RASMUSSEN, H. B., WERGE, T., HALL, H., TERENIUS, L. and JONSSON, E. G. (2008). Association between a disrupted-in-schizophrenia 1 (DISC1) single nucleotide polymorphism and schizophrenia in a combined Scandinavian case-control sample. *Schizophr. Res.* **106** 237–241.

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450

SHA, N., VANNUCCI, M., TADESSE, M. G., BROWN, P. J., DRAGONI, I., DAVIES, N., ROBERTS, T. C., CONTESTABILE, A., SALMON, M., BUCKLEY, C. and FALCIANI, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* **60** 812–828. MR2089459

SHAHBABA, B., SHACHAF, C. M. and YU, Z. (2012). A pathway analysis method for genome-wide association studies. *Stat. Med.* **31** 988–1000. MR2913874

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. MR3173712

SONNENBURG, S., RÄTSCH, G., SCHÄFER, C. and SCHÖLKOPF, B. (2006). Large scale multiple kernel learning. *J. Mach. Learn. Res.* **7** 1531–1565. MR2274416

STINGO, F. C., VANNUCCI, M. and DOWNEY, G. (2012). Bayesian wavelet-based curve classification via discriminant analysis with Markov random tree priors. *Statist. Sinica* **22** 465–488. MR2954348

STINGO, F. C., CHEN, Y. A., VANNUCCI, M., BARRIER, M. and MIRKES, P. E. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. *Ann. Appl. Stat.* **4** 2024–2048. MR2829945

STINGO, F. C., CHEN, Y. A., TADESSE, M. G. and VANNUCCI, M. (2011). Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Ann. Appl. Stat.* **5** 1978–2002. MR2884929

STINGO, F. C., GUINDANI, M., VANNUCCI, M. and CALHOUN, V. D. (2013). An integrative Bayesian modeling approach to imaging genetics. *J. Amer. Statist. Assoc.* **108** 876–891. MR3174670

SWARTZ, M. D., YU, R. K. and SHETE, S. (2008). Finding factors influencing risk: Comparing Bayesian stochastic search and standard variable selection methods applied to logistic regression models of cases and controls. *Stat. Med.* **27** 6158–6174. MR2522315

TZOURIO-MAZOYER, N., LANDEAU, B., PAPATHANASSIOU, D., CRIVELLO, F., ETARD, O., DELCROIX, N., MAZOYER, B. and JOLIOT, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15** 273–289.

VOUNOU, M., NICHOLS, T. E. and MONTANA, G. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage* **53** 1147–1159.

VOUNOU, M., JANOUSOVA, E., WOLZ, R., STEIN, J. L., THOMPSON, P. M., RUECKERT, D. and MONTANA, G. (2012). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *NeuroImage* **60** 700–716.

WALTZ, J. A., SCHWEITZER, J. B., GOLD, J. M., KURUP, P. K., ROSS, T. J., SALMERON, B. J., ROSE, E. J., MCCLURE, S. M. and STEIN, E. A. (2009). Patients with schizophrenia have a reduced neural response to both unpredictable and predictable primary reinforcers. *Neuropsychopharmacology* **34** 1567–1577.

WANG, H., NIE, F., HUANG, H., RISACHER, S. L., SAYKIN, A. J., SHEN, L. and THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2012a). Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics* **28** i127–i136.

WANG, H., NIE, F., HUANG, H., KIM, S., NHO, K., RISACHER, S. L., SAYKIN, A. J., SHEN, L. and THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2012b). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort. *Bioinformatics* **28** 229–237.

WEISS, K. M. (1989). Advantages of abandoning symptom-based diagnostic systems of research in schizophrenia. *Am. J. Orthopsychiatr.* **59** 324–330.

WU, L., CALHOUN, V. D., JUNG, R. E. and CAPRIHAN, A. (2015). Connectivity-based whole brain dual parcellation by group ICA reveals tract structures and decreased connectivity in schizophrenia. *Hum. Brain Mapp.* **36** 4681–4701.

XU, M.-Q., XING, Q.-H., ZHENG, Y.-L., LI, S., GAO, J.-J., HE, G., GUO, T.-W., FENG, G.-Y., XU, F. and HE, L. (2007). Association of AKT1 gene polymorphisms with risk of schizophrenia and with response to antipsychotics in the Chinese population. *J. Clin. Psychiatry* **68** 1358–1367.

YANG, H., LIU, J., SUI, J., PEARLSON, G. and CALHOUN, V. D. (2010). A hybrid machine learning method for fusing fMRI and genetic data: Combining both improves classification of schizophrenia. *Front. Human Neurosci.* **4** 1–9.

YU, Z., CHEN, J., SHI, H., STOEBER, G., TSANG, S.-Y. and XUE, H. (2006). Analysis of GABRB2 association with schizophrenia in German population with DNA sequencing and one-label extension method for SNP genotyping. *Clin. Biochem.* **39** 210–218.

ZHANG, L., GUINDANI, M. and VANNUCCI, M. (2015). Bayesian models for functional magnetic resonance imaging data analysis. *Wiley Interdiscip. Rev.: Comput. Stat.* **7** 21–41. MR3348719

ZHANG, Z., HUANG, H. and SHEN, D. (2014). Integrative analysis of multi-dimensional imaging genomics data for Alzheimer's disease prediction. *Front. Aging Neurosci.* **6** 1–9.

ZHANG, T., WIESEL, A. and GRECO, M. S. (2013). Multivariate generalized Gaussian distribution: Convexity and graphical models. *IEEE Trans. Signal Process.* **61** 4141–4148. MR3085302

ZHANG, H. H., AHN, J., LIN, X. and PARK, C. (2006). Gene selection using support vector machines with non-convex penalty. *Bioinformatics* **22** 88–95.

ZHU, H., KHONDKER, Z., LU, Z. and IBRAHIM, J. G. (2014). Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *J. Amer. Statist. Assoc.* **109** 977–990. MR3265670

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1572–1589
DOI: 10.1214/16-AOAS949

OPEN MODELS FOR REMOVAL DATA

BY ELENI MATECHOU, RACHEL S. MCCREA[1], BYRON J. T. MORGAN, DARRYN J. NASH AND RICHARD A. GRIFFITHS

University of Kent

Individuals of protected species, such as amphibians and reptiles, often need to be removed from sites before development commences. Usually, the population is considered to be closed. All individuals are assumed to (i) be present and available for detection at the start of the study period and (ii) remain at the site until the end of the study, unless they are detected. However, the assumption of population closure is not always valid. We present new removal models which allow for population renewal through birth and/or immigration, and population depletion through sampling as well as through death/emigration. When appropriate, productivity may be estimated and a Bayesian approach allows the estimation of the probability of total population depletion. We demonstrate the performance of the models using data on common lizards, *Zootoca vivipara*, and great crested newts, *Triturus cristatus*.

REFERENCES

AVERY, R. A. (1975). Clutch size and reproductive effort in the lizard *Lacerta vivipara* Jacquin. *Oecologia* **19** 165–170.

BARKER, R. J. (1997). Joint modeling of live recapture, tag-resight and tag-recovery data. *Biometrics* **53** 666–677.

BEEBEE, T. J. C. and GRIFFITHS, R. A. (2000). *Amphibians and Reptiles*. HarperCollins, London.

BESBEAS, P., FREEMAN, S. N., MORGAN, B. J. T. and CATCHPOLE, E. A. (2002). Integrating mark-recapture-recovery and census data to estimate animal abundance and demographic parameters. *Biometrics* **58** 540–547. MR1933532

BOHRMANN, T. F. and CHRISTMAN, M. C. (2013). Optimal allocation of sampling effort in depletion surveys. *J. Agric. Biol. Environ. Stat.* **18** 218–233. MR3067276

BROOKS, S. P., FREEMAN, S. N., GREENWOOD, J. J. D., KING, R. and MAZZETTA, C. (2008). Quantifying conservation concern—Bayesian statistics, birds and the red lists. *Biological Conservation* **141** 1436–1441.

COLE, D. J. and MCCREA, R. S. (2016). Parameter redundancy in discrete state-space and integrated models. *Biom. J.* To appear.

DEFRA (2015). https://www.gov.uk/guidance/reptiles-protection-surveys-and-licences. Natural England/Defra.

DORAZIO, R. M., JELKS, H. L. and JORDAN, F. (2005). Improving removal-based estimates of abundance by sampling a population of spatially distinct subpopulations. *Biometrics* **61** 1093–1101. MR2216203

DORAZIO, R. M., MUKHERJEE, B., ZHANG, L., GHOSH, M., JELKS, H. L. and JORDAN, F. (2008). Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics* **64** 635–644. MR2432438

Key words and phrases. Common lizard, depletion, great crested newts, RJMCMC, stopover model.

GENT, T. and GIBSON, S. E. (1998). *Herpetofauna Workers' Manual*. Joint Nature Conservation Committee, Peterborough.

GERMANO, J. M., FIELD, K. J., GRIFFITHS, R. A., CLULOW, S., FOSTER, J., HARDING, G. and SWAISGOOD, R. R. (2015). Mitigation-driven translocations: Are we moving wildlife in the right direction? *Frontiers in Ecology and the Environment* **13** 100–105.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

HGBI (1998). Herpetofauna Groups of Britain and Ireland—Evaluating Local Mitigation/translocation Programmes: Maintaining Best Practice and Lawful Standards. HGBI Advisory Notes for Amphibian and Reptile groups. Preprint.

JOPPA, L. N., WILLIAMS, K. R., TEMPLE, S. A. and CASPER, G. S. (2009). Environmental factors affecting sampling success of artifical cover objects. *Herpetological Conservation and Biology* **5** 143–148.

KENDALL, W. L. and NICHOLS, J. D. (2002). Estimating state-transition probabilities for unobservable states using capture-recapture/resighting data. *Ecology* **83** 3276–3284.

KENDALL, W. L., NICHOLS, J. D. and HINES, J. E. (1997). Estimating temporary emigration using capture-recapture data with Pollock's robust design. *Ecology* **78** 563–578.

LEWIS, B., GRIFFITHS, R. A., WILKINSON, J. W. and ARNELL, A. (2014). Examining the Fate of Local Great Crested Newt Populations Following Licensed Developments. Report WM031, London: Department for Environment, Food and Rural Affairs.

MATECHOU, E., DENNIS, E. B., FREEMAN, S. N. and BRERETON, T. (2014). Monitoring abundance and phenology in (multivoltine) butterfly species: A novel mixture model. *Journal of Applied Ecology* **51** 766–775.

MATECHOU, E., NICHOLLS, G., MORGAN, B. J. T., COLLAZO, J. A. and LYONS, J. E. (2015). Bayesian analysis of Jolly–Seber type models; incorporating heterogeneity in arrival and departure. *Environ. Ecol. Stat.* To appear. DOI:10.1007/s10651-016-0352-0.

MATECHOU, E., MCCREA, R. S., MORGAN, B. J. T., NASH, J. D. and GRIFFITHS, R. A. (2016). Supplement to "Open models for removal data." DOI:10.1214/16-AOAS949SUPP.

MCCREA, R. S. and MORGAN, B. J. T. (2014). *Analysis of Capture–Recapture Data*. Chapman & Hall/CRC, Boca Raton, FL. MR3330977

MORAN, P. A. P. (1951). A mathematical theory of animal trapping. *Biometrika* **38** 307–311.

MORGAN, B. J. T. (2009). *Applied Stochastic Modelling*, 2nd ed. CRC Press, Boca Raton, FL. MR2468192

PLEDGER, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* **56** 434–442.

PRADEL, R. (2005). Multievent: An extension of multistate capture-recapture models to uncertain states. *Biometrics* **61** 442–447. MR2140915

READING, C. J. (1997). A proposed standard method for surveying reptiles on dry lowland heath. *Journal of Applied Ecology* **34** 1057–1069.

RIDOUT, M. S. and MORGAN, B. J. T. (1991). Modelling digit preference in fecundability studies. *Biometrics* **47** 1423–1434.

RUIZ, P. and LAPLANCHE, C. A. (2010). A hierarchical model to estimate the abundance and biomass of salmonids by using removal sampling and biometric data from multiple locations. *Canadian Journal of Fisheries and Aquatic Sciences* **67** 2032–2044.

SEWELL, D., BEEBEE, T. J. C. and GRIFFITHS, R. A. (2010). Optimising biodiversity assessments by volunteers: The application of occupancy modelling to large-scale amphibian surveys. *Biological Conservation* **143** 2102–2110.

SEWELL, D., GUILLERA-ARROITA, G., GRIFFITHS, R. A. and BEEBEE, T. J. C. (2012). When is a species declining? Optimizing survey effort to detect population changes in reptiles. *PLOS ONE* **7** e43387.

VAN DAMME, R., BAUWENS, D. and VERHEYEN, R. F. (1990). Evolutionary rigidity of thermal physiology—The case of the cool temperate lizard, *Lacerta vivipara*. *Oikos* **57** 61–67.

ZIPPIN, C. (1956). An evaluation of the removal method of estimating animal populations. *Biometrics* **12** 163–189.

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1590–1618
DOI: 10.1214/16-AOAS950

SPATIO-TEMPORAL ASSIMILATION OF MODELLED CATCHMENT LOADS WITH MONITORING DATA IN THE GREAT BARRIER REEF

BY DANIEL W. GLADISH[*], PETRA M. KUHNERT[*,1], DANIEL E. PAGENDAM[*], CHRISTOPHER K. WIKLE[†], REBECCA BARTLEY[*], ROSS D. SEARLE[*], ROBIN J. ELLIS[‡], CAMERON DOUGALL[§], RYAN D. R. TURNER[‡], STEPHEN E. LEWIS[¶], ZOË T. BAINBRIDGE[¶] AND JON E. BRODIE[¶]

Commonwealth Scientific and Industrial Research Organisation (CSIRO)[*], University of Missouri[†], Department of Science Information Technology and Innovation[‡], Department of Natural Resources and Mines[§] and James Cook University[¶]

Soil erosion and sediment transport into waterways and the ocean can adversely affect water clarity, leading to the deterioration of marine ecosystems such as the iconic Great Barrier Reef (GBR) in Australia. Quantifying a sediment load and its associated uncertainty is an important task in delineating how changes in management practices can contribute to improvements in water quality, and therefore continued sustainability of the GBR. However, monitoring data are spatially (and often temporally) sparse, making load estimation complicated, particularly when there are lengthy periods between sampling or during peak flow periods of major events when samples cannot be safely taken.

We develop a spatio-temporal statistical model that is mechanistically motivated by a process-based deterministic model called Dynamic SedNet. The model is developed within a Bayesian hierarchical modelling framework that uses dimension reduction to accommodate seasonal and spatial patterns to assimilate monitored sediment concentration and flow data with output from Dynamic SedNet. The approach is applied in the Upper Burdekin catchment in Queensland, Australia, where we obtain daily estimates of sediment concentrations, stream discharge volumes and sediment loads at 411 spatial locations across 20 years. Our approach provides a method for assimilating both monitoring data and modelled output, providing a statistically rigorous method for quantifying uncertainty through space and time that was previously unavailable through process-based models.

REFERENCES

AKSOY, H. and KAVVAS, M. L. (2005). A review of hillslope and watershed scale erosion and sediment transport models. *Catena* **64** 247–271.

ARMOUR, J. D., HATELEY, L. R. and PITT, G. L. (2009). Catchment modelling of sediment, nitrogen and phosphorus nutrient loads with SedNet/ANNEX in the Tully–Murray basin. *Marine and Freshwater Research* **60** 1091–1096.

Key words and phrases. Water quality, Bayesian hierarchical model, SedNet, catchment modelling, spatio-temporal.

BAINBRIDGE, Z. T., LEWIS, S. E., SMITHERS, S. G., KUHNERT, P. M., HENDERSON, B. L. and BRODIE, J. E. (2014). Fine-suspended sediment and water budgets for a large, seasonally dry tropical catchment: Burdekin river catchment, Queensland, Australia. *Water Resources Research* **50** 9067–9087.

BARTLEY, R., WILKINSON, S. N., HAWDON, A., ABBOTT, B. N. and POST, D. A. (2010). Impacts of improved grazing land management on sediment yields. Part 2: Catchment response. *Journal of Hydrology* **389** 249–259.

BARTLEY, R., BAINBRIDGE, Z. T., LEWIS, S. E., KROON, F. J., WILKINSON, S. N., BRODIE, J. E. and SILBURN, D. M. (2014). Relating sediment impacts on coral reefs to watershed sources, processes and management: A review. *Sci. Total Environ.* **468–469** 1138–1153.

BERLINER, L. M. (1996). Hierarchical Bayesian time series models. In *Maximum Entropy and Bayesian Methods* (*Santa Fe*, *NM*, 1995) (K. M. Hanson and R. N. Silver, eds.). *Fund. Theories Phys.* **79** 15–22. Kluwer Academic, Dordrecht. MR1446713

BEVEN, K. and FREER, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology* **249** 11–29.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc., Ser. B Stat. Methodol.* **26** 211–252.

BRODIE, J., SCHROEDER, T., ROHDE, K., FAITHFUL, J., MASTERS, B., DEKKER, A., BRANDO, V. and MAUGHAN, M. (2010). Dispersal of suspended sediments and nutrients in the Great Barrier Reef lagoon during river-discharge events: Conclusions from satellite remote sensing and concurrent flood-plume sampling. *Marine and Freshwater Research* **61** 651–664.

BRODIE, J. E., DEVLIN, M., HAYNES, D. and WATERHOUSE, J. (2011). Assessment of the eutrophication status of the Great Barrier Reef lagoon (Australia). *Biogeochemistry* **106** 281–302.

BRODIE, J. E., KROON, F. J., SCHAFFELKE, B., WOLANSKI, E. C., LEWIS, S. E., DEVLIN, M. J., BOHNET, I. C., BAINBRIDGE, Z. T., WATERHOUSE, J. and DAVIS, A. M. (2012). Terrestrial pollutant runoff to the Great Barrier Reef: An update of issues, priorities and management responses. *Mar. Pollut. Bull.* **65** 81–100.

BRYNJARSDÓTTIR, J. and BERLINER, L. M. (2014). Dimension-reduced modeling of spatio-temporal processes. *J. Amer. Statist. Assoc.* **109** 1647–1659. MR3293617

CHEN, D., DAHLGREN, R. A., SHEN, Y. and LU, J. (2012). A Bayesian approach for calculating variable total maximum daily loads and uncertainty assessment. *Science of the Total Environment* **430** 59–67.

CHIEW, F. H. S., PEEL, M. C. and WESTERN, A. W. (2002). Application and testing of the simple rainfall-runoff model SIMHYD. In *Mathematical Models of Small Watershed Hydrology and Applications* (V. P. Singh and D. K. Frevert , eds.) 335–367. Water Resources Publications, Littleton, CO.

CLARK, J. S. (2005). Why environmental scientists are becoming bayesians. *Ecology Letters* **8** 2–14.

COHN, T. A. (1995). Recent advances in statistical methods for the estimation of sediment and nutrient transport in rivers. *Reviews of Geophysics* **33** 1117–1123.

COHN, T. A., CAULDER, D. L., GILROY, E. J., ZYNJUK, L. D. and SUMMERS, R. M. (1992). The validity of a simple statistical model for estimating fluvial constituent loads: An empirical study involving nutrient loads entering Chesapeake Bay. *Water Resources Research* **28** 2353–2363.

COOPER, D. M. and WATTS, C. D. (2002). A comparison of river load estimation techniques: Application to dissolved organic carbon. *Environmetrics* **13** 733–750.

CRAWFORD, C. G. (1991). Estimation of suspended-sediment rating curves and mean suspended-sediment loads. *Journal of Hydrology* **129** 331–348.

CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ. MR2848400

DE'ATH, G., FABRICIUS, K. E., SWEATMAN, H. and PUOTINEN, M. (2012). The 27-year decline of coral cover on the Great Barrier Reef and its causes. *Proc. Natl. Acad. Sci. USA* **109** 17995–17999.

DOHERTY, J. and HUNT, R. J. (2009). Two statistics for evaluating parameter identifiability and error reduction. *Journal of Hydrology* **366** 119–127.

DOHERTY, J. and JOHNSTON, J. M. (2003). Methodologies for calibration and predictive analysis of a watershed model. *Journal of the American Water Resources Association* **39** 251–265.

FABRICIUS, K. E., LOGAN, M., WEEKS, S. and BRODIE, J. (2014). The effects of river run-off on water clarity across the central Great Barrier Reef. *Mar. Pollut. Bull.* **84** 191–200.

FURNAS, M. J. (2003). Catchments and corals: Terrestrial runoff to the Great Barrier Reef, report, Australian Institute of Marine Science and CRC Reef, Townsville, Australia.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6** 721–741.

GLADISH, D. W. and WIKLE, C. K. (2014). Physically motivated scale interaction parameterization in reduced rank quadratic nonlinear dynamic spatio-temporal models. *Environmetrics* **25** 230–244. MR3256458

GLADISH, D. W., KUHNERT, P. M., PAGENDAM, D. E., WIKLE, C. K., BARTLEY, R. B., SEARLE, R. D., ELLIS, R. J., DOUGALL, C., TURNER, R. D. R., LEWIS, S. E., BAINBRIDGE, Z. T. and BRODIE, J. E. (2016). Supplement to "Spatio-temporal assimilation of modelled catchment loads with monitoring data in the Great Barrier Reef." DOI:10.1214/16-AOAS950SUPP.

GRAF, A., BOGENA, H. R., DRÜE, C., HARDELAUF, H., PÜTZ, T., HEINEMANN, G. and VEREECKEN, H. (2014). Spatiotemporal relations between water budget components and soil water content in a forested tributary catchment. *Water Resources Research* **50** 4837–4857.

KUHNERT, P. M., HENDERSON, B. L., LEWIS, S. E., BAINBRIDGE, Z. T., WILKINSON, S. N. and BRODIE, J. E. (2012). Quantifying total suspended sediment export from the Burdekin River catchment using the loads regression estimator tool. *Water Resources Research* **48** W04533.

LETCHER, R. A., JAKEMAN, A. J., CALFAS, M., LINFORTH, S., BAGINSKA, B. and LAWRENCE, I. (2002). A comparison of catchment water quality models and direct estimation techniques. *Environmental Modelling and Software* **17** 77–85.

LE COZ, J., RENARD, B., BONNIFAIT, L., BRANGER, F. and LE BOURSICAUD, R. (2014). Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: A Bayesian approach. *Journal of Hydrology* **509** 573–587.

LITTLEWOOD, I. G. and MARSH, T. J. (2005). Annual freshwater river mass loads from Great Britain, 1975–1994: Estimation algorithm, database and monitoring network issues. *Journal of Hydrology* **304** 221–237.

LIU, Y., YANG, P., HU, C. and GUO, H. (2008). Water quality modeling for load reduction under uncertainty: A Bayesian approach. *Water Research* **42** 3305–3314.

LOUGH, J. M., LEWIS, S. E. and CANTIN, N. E. (2015). Freshwater impacts in the central Great Barrier Reef: 1648-2011. *Coral Reefs* **34** 1–13.

MANTOVAN, P. and TODINI, E. (2006). Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. *Journal of Hydrology* **330** 368–381.

MCCULLOCH, M., FALLON, S., WYNDHAM, T., HENDY, E., LOUGH, J. and BARNES, D. (2003). Coral record of increased sediment flux to the inner Great Barrier Reef since European settlement. *Nature* **421** 727–730.

MOYEED, R. A. and CLARKE, R. T. (2005). The use of Bayesian methods for fitting rating curves, with case studies. *Advances in Water Resources* **28** 807–818.

OLESON, J. J. and WIKLE, C. K. (2013). Predicting infectious disease outbreak risk via migratory waterfowl vectors. *J. Appl. Stat.* **40** 656–673. MR3047308

PAGENDAM, D. E., KUHNERT, P. M., LEEDS, W. B., WIKLE, C. K., BARTLEY, R. and PETERSON, E. E. (2014). Assimilating catchment processes with monitoring data to estimate sediment loads to the Great Barrier Reef. *Environmetrics* **25** 214–229. MR3256457

PERRIN, C., MICHEL, C. and ANDRÉASSIAN, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology* **279** 275–289.

R CORE TEAM (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

REEF WATER QUALITY PROTECTION PLAN SECRETARIAT (2013). Reef Water Quality Protection Plan. Available at http://www.reefplan.qld.gov.au/resources/assets/reef-plan-2013.pdf.

REITAN, T. and PETERSEN-ØVERLEIR, A. (2011). Dynamic rating curve assessment in unstable rivers using Ornstein–Uhlenbeck processes. *Water Resources Research* **47** W02524.

RENARD, K. G., FOSTER, G. R., WEESIES, G. A., MCCOOL, D. K. and YODER, D. C. (1997). Predicting soil erosion by water: A guide to conservation planning with the revised universal soil loss equation (RUSLE). U.S. Department of Agriculture, Agriculture Handbook No. 703, 404 pp.

ROSEWELL, C. J. (1993). SOILLOSS—A program to assis in the selection of management practices to reduce erosion. Technical Report No. Technical Handbook No.11, 2nd edition, Soil Conservation Service, NSW.

RUSTOMJI, P. and WILKINSON, S. N. (2008). Applying bootstrap resampling to quantify uncertainty in fluvial suspended sediment loads estimated using rating curves. *Water Resources Research* **44** W09435.

SALAZAR, E., SANSÓ, B., FINLEY, A. O., HAMMERLING, D., STEINSLAND, I., WANG, X. and DELAMATER, P. (2011). Comparing and blending regional climate model predictions for the American Southwest. *J. Agric. Biol. Environ. Stat.* **16** 586–605. MR2862300

SANSÓ, B. and GUENNI, L. (1999). Venezuelan rainfall data analysed by using a Bayesian space–time model. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **48** 345–362.

SCHMELTER, M. L., HOOTEN, M. B. and STEVENS, D. K. (2011). Bayesian sediment transport model for unisize bed load. *Water Resources Research* **47** W11514.

TOMKINS, K. M. (2014). Uncertainty in streamflow rating curves: Methods, controls and consequences. *Hydrological Processes* **28** 464–481.

WALLING, D. E. and WEBB, B. W. (1985). Estimating the discharge of contaminants to coastal waters by rivers: Some cautionary comments. *Mar. Pollut. Bull.* **16** 488–492.

WANG, Y. G., KUHNERT, P. and HENDERSON, B. (2011). Load estimation with uncertainties from opportunistic sampling data—A semiparametric approach. *Journal of Hydrology* **396** 148–157.

WIKLE, C. K. and BERLINER, L. M. (2007). A Bayesian tutorial for data assimilation. *Phys. D* **230** 1–16. MR2345198

WIKLE, C. K. and HOOTEN, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. *TEST* **19** 417–451. MR2745992

WILKINSON, S. N., HANCOCK, G. J., BARTLEY, R., HAWDON, A. A. and KEEN, R. J. (2013). Using sediment tracing to assess processes and spatial patterns of erosion in grazed rangelands, burdekin river basin, Australia. *Agriculture, Ecosystems and Environment* **180** 90–102.

WILKINSON, S. N., DOUGALL, C., KINSEY-HENDERSON, A. E., SEARLE, R. D., ELLIS, R. J. and BARTLEY, R. (2014). Development of a time-stepping sediment budget model for assessing land use impacts in large river basins. *Sci. Total Environ.* **468–469** 1210–1224.

WU, W., CLARK, J. S. and VOSE, J. M. (2010). Assimilating multi-source uncertainties of a parsimonious conceptual hydrological model using hierarchical Bayesian modeling. *Journal of Hydrology* **394** 436–446.

WU, G., HOLAN, S. H. and WIKLE, C. K. (2013). Hierarchical Bayesian spatio-temporal Conway-Maxwell Poisson models with dynamic dispersion. *J. Agric. Biol. Environ. Stat.* **18** 335–356. MR3110897

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1619–1638
DOI: 10.1214/16-AOAS952

MOLECULAR QTL DISCOVERY INCORPORATING GENOMIC ANNOTATIONS USING BAYESIAN FALSE DISCOVERY RATE CONTROL[1]

BY XIAOQUAN WEN

University of Michigan

Mapping molecular QTLs has emerged as an important tool for understanding the genetic basis of cell functions. With the increasing availability of functional genomic data, it is natural to incorporate genomic annotations into QTL discovery. Discovering molecular QTLs is typically framed as a multiple hypothesis testing problem and solved using false discovery rate (FDR) control procedures. Currently, most existing statistical approaches rely on obtaining p-values for each candidate locus through permutation-based schemes, which are not only inconvenient for incorporating highly informative genomic annotations but also computationally inefficient. In this paper, we discuss a novel statistical approach for integrative QTL discovery based on the theoretical framework of Bayesian FDR control. We use a Bayesian hierarchical model to naturally integrate genomic annotations into molecular QTL mapping and propose an empirical Bayes-based computational procedure to approximate the necessary posterior probabilities to achieve high computational efficiency. Through theoretical arguments and simulation studies, we demonstrate that the proposed approach rigorously controls the desired type I error rate and greatly improves the power of QTL discovery when incorporating informative annotations. Finally, we demonstrate our approach by analyzing the expression-genotype data from 44 human tissues generated by the GTEx project. By integrating the simple annotation of SNP distance to transcription start sites, we discover more genes that harbor expression-associated SNPs in all 44 tissues, with an average increase of 1485 genes per tissue.

REFERENCES

ARDLIE, K. G., DELUCA, D. S., SEGRÈ, A. V., SULLIVAN, T. J., YOUNG, T. R., GELFAND, E. T., TROWBRIDGE, C. A., MALLER, J. B., TUKIAINEN, T., LEK, M. et al. (2015). The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348** 648–660.

BALL, R. D. (2001). Bayesian methods for quantitative trait loci mapping based on model selection: Approximate analysis using the Bayesian information criterion. *Genetics* **159** 1351–1364.

BANOVICH, N. E., LAN, X., MCVICKER, G., VAN DE GEIJN, B., DEGNER, J. F., BLISCHAK, J. D., ROUX, J., PRITCHARD, J. K. and GILAD, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet*. **10** e1004663.

Key words and phrases. Molecular QTL, genomic annotations, Bayesian FDR control, QTL mapping.

BARREIRO, L. B., TAILLEUX, L., PAI, A. A., GICQUEL, B., MARIONI, J. C. and GILAD, Y. (2012). Deciphering the genetic architecture of variation in the immune response to mycobacterium tuberculosis infection. *Proc. Natl. Acad. Sci. USA* **109** 1204–1209.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **57** 289–300. MR1325392

BERISA, T. and PICKRELL, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32** 283–285.

BREITLING, R., LI, Y., TESSON, B. M., FU, J., WU, C., WILTSHIRE, T., GERRITS, A., BYSTRYKH, L. V., DE HAAN, G., SU, A. I. et al. (2008). Genetical genomics: Spotlight on QTL hotspots. *PLoS Genet.* **4** e1000232.

CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456.

CHURCHILL, G. A. and DOERGE, R. W. (1994). Empirical threshold values for quantitative trait mapping *Genetics* **138** 963–971.

DEGNER, J. F., PAI, A. A., PIQUE-REGI, R., VEYRIERAS, J.-B., GAFFNEY, D. J., PICKRELL, J. K., DE LEON, S., MICHELINI, K., LEWELLEN, N., CRAWFORD, G. E. et al. (2012). DNase [thinsp] I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482** 390–394.

DE LA CRUZ, O., WEN, X., KE, B., SONG, M. and NICOLAE, D. L. (2010). Gene, region and pathway level analyses in whole-genome studies. *Genet. Epidemiol.* **34** 222–231.

DING, Z., NI, Y., TIMMER, S. W., LEE, B.-K., BATTENHOUSE, A., LOUZADA, S., YANG, F., DUNHAM, I., CRAWFORD, G. E., LIEB, J. D. et al. (2014). Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLOS Genet.* **10** e1004798.

DOERGE, R. W. and CHURCHILL, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142** 285–294.

ENCODE PROJECT CONSORTIUM et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74.

FLUTRE, T., WEN, X., PRITCHARD, J. and STEPHENS, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* **9** e1003486.

KUNDAJE, A., MEULEMAN, W., ERNST, J., BILENKY, M., YEN, A., HERAVI-MOUSSAVI, A., KHERADPOUR, P., ZHANG, Z., WANG, J., ZILLER, M. J. et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* **518** 317–330.

LAPPALAINEN, T., SAMMETH, M., FRIEDLÄNDER, M. R., HOEN, P. A. C. T., MONLONG, J., RIVAS, M. A., GONZÀLEZ-PORTA, M., KURBATOVA, N., GRIEBEL, T., FERREIRA, P. G. et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501** 506–511.

LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3** e161.

LEVINE, R. A. and CASELLA, G. (2001). Implementations of the Monte Carlo EM algorithm. *J. Comput. Graph. Statist.* **10** 422–439. MR1939033

MARANVILLE, J. C., LUCA, F., RICHARDS, A. L., WEN, X., WITONSKY, D. B., BAXTER, S., STEPHENS, M., DI RIENZO, A. and GIBSON, G. (2011). Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes. *PLOS Genet.* **7** e1002162.

MARIN, J.-M. and ROBERT, C. P. (2007). *Bayesian Core*: *A Practical Approach to Computational Bayesian Statistics*. Springer, New York. MR2289769

MCVICKER, G., VAN DE GEIJN, B., DEGNER, J. F., CAIN, C. E., BANOVICH, N. E., RAJ, A., LEWELLEN, N., MYRTHIL, M., GILAD, Y. and PRITCHARD, J. K. (2013). Identification of genetic variants that affect histone modifications in human cells. *Science* **342** 747–749.

MÜLLER, P., PARMIGIANI, G., ROBERT, C. and ROUSSEAU, J. (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *J. Amer. Statist. Assoc.* **99** 990–1001. MR2109489

NETO, E. C., KELLER, M. P., BROMAN, A. F., ATTIE, A. D., JANSEN, R. C., BROMAN, K. W. and YANDELL, B. S. (2012). Quantile-based permutation thresholds for quantitative trait loci hotspots. *Genetics* **191** 1355–1365.

NETO, E. C., BROMAN, A. T., KELLER, M. P., ATTIE, A. D., ZHANG, B., ZHU, J. and YANDELL, B. S. (2013). Modeling causality for pairs of phenotypes in system genetics. *Genetics* **193** 1003–1013.

NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.

PIQUE-REGI, R., DEGNER, J. F., PAI, A. A., GAFFNEY, D. J., GILAD, Y. and PRITCHARD, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21** 447–455.

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

SHABALIN, A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28** 1353–1358.

SILLANPÄÄ, M. J. and ARJAS, E. (1999). Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* **151** 1605–1619.

STEGLE, O., PARTS, L., PIIPARI, M., WINN, J. and DURBIN, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7** 500–507.

STEPHENS, D. A. and FISCH, R. D. (1998). Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics* **54** 1334–1347.

STOREY, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035. MR2036398

SUL, J. H., RAJ, T., DE JONG, S., DE BAKKER, P. I., RAYCHAUDHURI, S., OPHOFF, R. A., STRANGER, B. E., ESKIN, E. and HAN, B. (2015). Accurate and fast multiple-testing correction in eQTL studies. *The American Journal of Human Genetics* **96** 857–868.

VEYRIERAS, J.-B., KUDARAVALLI, S., KIM, S. Y., DERMITZAKIS, E. T., GILAD, Y., STEPHENS, M. and PRITCHARD, J. K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLOS Genet.* **4** e1000214.

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

WEN, X. (2011). Bayesian analysis of genetic association data, accounting for heterogeneity. Ph.D. thesis, Univ. Chicago.

WEN, X. (2014). Bayesian model selection in complex linear systems, as illustrated in genetic association studies. *Biometrics* **70** 73–83. MR3251668

WEN, X. (2015). Bayesian model comparison in genetic association analysis: Linear mixed modeling and SNP set testing. *Biostatistics* **16** 701–712. MR3449837

WEN, X. (2016). Supplement to "Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control." DOI:10.1214/16-AOAS952SUPP.

WEN, X., LUCA, F. and PIQUE-REGI, R. (2015). Cross-population joint analysis of eQTLs: Fine mapping and functional annotation. *PLoS Genet.* **11** e1005176.

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

YI, N., YANDELL, B. S., CHURCHILL, G. A., ALLISON, D. B., EISEN, E. J. and POMP, D. (2005). Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics* **170** 1333–1344.

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1639–1672
DOI: 10.1214/16-AOAS953

MORTALITY AND LIFE EXPECTANCY FORECASTING FOR A GROUP OF POPULATIONS IN DEVELOPED COUNTRIES: A MULTILEVEL FUNCTIONAL DATA METHOD

BY HAN LIN SHANG

Australian National University

A multilevel functional data method is adapted for forecasting age-specific mortality for two or more populations in developed countries with high-quality vital registration systems. It uses multilevel functional principal component analysis of aggregate and population-specific data to extract the common trend and population-specific residual trend among populations. If the forecasts of population-specific residual trends do not show a long-term trend, then convergence in forecasts may be achieved. This method is first applied to age- and sex-specific data for the United Kingdom, and its forecast accuracy is then further compared with several existing methods, including independent functional data and product-ratio methods, through a multi-country comparison. The proposed method is also demonstrated by age-, sex- and state-specific data in Australia, where the convergence in forecasts can possibly be achieved by sex and state. For forecasting age-specific mortality, the multilevel functional data method is more accurate than the other coherent methods considered. For forecasting female life expectancy at birth, the multilevel functional data method is outperformed by the Bayesian method of Raftery, Lalic and Gerland [*Demogr. Res.* **30** (2014) 795–822]. For forecasting male life expectancy at birth, the multilevel functional data method performs better than the Bayesian methods in terms of point forecasts, but less well in terms of interval forecasts. Supplementary materials for this article are available online.

REFERENCES

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

ALKEMA, L., RAFTERY, A. E., GERLAND, P., CLARK, S. J., PELLETIER, F., BUETTNER, T. and HEILIG, G. K. (2011). Probabilistic projections of the total fertility rate for all countries. *Demography* **48** 815–839.

AUE, A., NORINHO, D. D. and HÖRMANN, S. (2015). On the prediction of stationary functional time series. *J. Amer. Statist. Assoc.* **110** 378–392. MR3338510

BIATAT, V. D. and CURRIE, I. D. (2010). Joint models for classification and comparison of mortality in different countries. In *Proceedings of* 25*th International Workshop on Statistical Modelling* (A. W. Bowman, ed.) 89–94. Glasgow.

BOOTH, H. (2006). Demographic forecasting: 1980–2005 in review. *International Journal of Forecasting* **22** 547–581.

Key words and phrases. Augmented common factor method, coherent forecasts, functional time series, life expectancy forecasting, mortality forecasting, product-ratio method.

BOOTH, H., MAINDONALD, J. and SMITH, L. (2002). Applying Lee–Carter under conditions of variable mortality decline. *Popul. Stud. (Camb.)* **56** 325–336.
BOOTH, H. and TICKLE, L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science* **3** 3–43.
BOX, G. E. P., JENKINS, G. M. and REINSEL, G. C. (2008). *Time Series Analysis: Forecasting and Control*, 4th ed. Wiley, Hoboken, NJ. MR2419724
CAIRNS, A. J. G., BLAKE, D., DOWD, K., COUGHLAN, G. D., EPSTEIN, D. and KHALAF-ALLAH, M. (2011a). Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance Math. Econom.* **48** 355–367. MR2820048
CAIRNS, A. J. G., BLAKE, D., DOWD, K., COUGHLAN, G. D. and KHALAF-ALLAH, M. (2011b). Bayesian stochastic mortality modelling for two populations. *Astin Bull.* **41** 29–59. MR2828982
CHERNICK, M. R. (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers*, 2nd ed. Wiley, Hoboken, NJ. MR2355547
CHIOU, J.-M. (2012). Dynamical functional prediction and classification, with application to traffic flow prediction. *Ann. Appl. Stat.* **6** 1588–1614. MR3058676
CRAINICEANU, C. M. and GOLDSMITH, J. A. (2010). Bayesian functional data analysis using WinBUGS. *Journal of Statistical Software* **32**. DOI:10.18637/jss.v032.i11.
CRAINICEANU, C. M., STAICU, A.-M. and DI, C.-Z. (2009). Generalized multilevel functional regression. *J. Amer. Statist. Assoc.* **104** 1550–1561. MR2750578
CUESTA-ALBERTOS, J. A. and FEBRERO-BANDE, M. (2010). A simple multiway ANOVA for functional data. *TEST* **19** 537–557. MR2746001
CURRIE, I. D., DURBAN, M. and EILERS, P. H. C. (2004). Smoothing and forecasting mortality rates. *Stat. Model.* **4** 279–298. MR2086492
DELWARDE, A., DENUIT, M., GUILLÉN, M. and VIDIELLA-I-ANGUERA, A. (2006). Application of the Poisson log-bilinear projection model to the G5 mortality experience. *Belg. Actuar. Bull.* **6** 54–68.
DI, C.-Z., CRAINICEANU, C. M., CAFFO, B. S. and PUNJABI, N. M. (2009). Multilevel functional principal component analysis. *Ann. Appl. Stat.* **3** 458–488. MR2668715
DOWD, K., CAIRNS, A. J. G., BLAKE, D., COUGHLAN, G. D. and KHALAF-ALLAH, M. (2011). A gravity model of mortality rates for two related populations. *N. Am. Actuar. J.* **15** 334–356. MR2835504
GIROSI, F. and KING, G. (2008). *Demographic Forecasting*. Princeton Univ. Press, Princeton, NJ.
GNEITING, T. and KATZFUSS, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Applications* **1** 125–151.
GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548
GREVEN, S., CRAINICEANU, C., CAFFO, B. and REICH, D. (2010). Longitudinal functional principal component analysis. *Electron. J. Stat.* **4** 1022–1054. MR2727452
HALL, P. and VIAL, C. (2006). Assessing the finite dimensionality of functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 689–705. MR2301015
HE, X. and NG, P. (1999). COBS: Qualitatively constrained smoothing via linear programming. *Comput. Statist.* **14** 315–337.
HOFF, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer, New York. MR2648134
HUMAN MORTALITY DATABASE (2015). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Accessed at 8 March 2013. Available at http://www.mortality.org.
HYNDMAN, R. J. (2010). addb: Australian Demographic Data Bank. R package version 3.223. Available at http://robjhyndman.com/software/addb/.
HYNDMAN, R. J., BOOTH, H. and YASMEEN, F. (2013). Coherent mortality forecasting: The product-ratio method with functional time series models. *Demography* **50** 261–283.

HYNDMAN, R. J. and KHANDAKAR, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software* **27**. DOI:10.18637/jss.v027.i03.

HYNDMAN, R. J. and ULLAH, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Comput. Statist. Data Anal.* **51** 4942–4956. MR2364551

HYNDMAN, R. J. and SHANG, H. L. (2009). Forecasting functional time series. *J. Korean Statist. Soc.* **38** 199–211. MR2750314

HYNDMAN, R. J., AHMED, R. A., ATHANASOPOULOS, G. and SHANG, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Comput. Statist. Data Anal.* **55** 2579–2589. MR2802337

JANSSEN, F., VAN WISSEN, L. J. G. and KUNST, A. E. (2013). Including the smoking epidemic in internationally coherent mortality projection. *Demography* **50** 1341–1362.

JARNER, S. F. and KRYGER, E. M. (2011). Modelling adult mortality in small populations: The saint model. *Astin Bull.* **41** 377–418. MR2858780

KWIATKOWSKI, D., PHILLIPS, P. C. B., SCHMIDT, P. and SHIN, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *J. Econometrics* **54** 159–178.

LEE, R. (2000). The Lee–Carter method for forecasting mortality, with various extensions and applications. *N. Am. Actuar. J.* **4** 80–93. MR2011262

LEE, R. D. (2006). Mortality forecasts and linear life expectancy trends. In *Perspectives on Mortality Forecasting. Vol. III. The Linear Rise in Life Expectancy*: *History and Prospects* (T. Bengtsson, ed.). *Social Insurance Studies* **3** 19–39. Swedish National Social Insurance Board, Stockholm.

LEE, R. D. and CARTER, L. R. (1992). Modeling and forecasting U.S. mortality. *J. Amer. Statist. Assoc.* **87** 659–671.

LEE, R. D. and MILLER, T. (2001). Evaluating the performance of the Lee–Carter method for forecasting mortality. *Demography* **38** 537–549.

LI, J. (2013). A Poisson common factor model for projecting mortality and life expectancy jointly for females and males. *Popul. Stud.* (*Camb.*) **67** 111–126.

LI, J. S.-H. and HARDY, M. R. (2011). Measuring basis risk in longevity hedges. *N. Am. Actuar. J.* **15** 177–200. MR2835496

LI, N. and LEE, R. (2005). Coherent mortality forecasts for a group of population: An extension of the Lee–Carter method. *Demography* **42** 575–594.

LI, N., LEE, R. and GERLAND, P. (2013). Extending the Lee–Carter method to model the rotation of age patterns of mortality decline for long-term projections. *Demography* **50** 2037–2051.

MORRIS, J. S. and CARROLL, R. J. (2006). Wavelet-based functional mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 179–199. MR2188981

MORRIS, J. S., VANNUCCI, M., BROWN, P. J. and CARROLL, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *J. Amer. Statist. Assoc.* **98** 573–583. MR2011673

OEPPEN, J. and VAUPEL, J. W. (2002). Demography. Broken limits to life expectancy. *Science* **296** 1029–1031.

PAMPEL, F. C. (2005). Forecasting sex differences in mortality from lung cancer in high-income nations: The contribution of smoking. *Demogr. Res.* **13** 455–484.

PRESTON, S. H., HEUVELINE, P. and GUILLOT, M. (2001). *Demography*: *Measuring and Modelling Population Process*. Blackwell, Oxford, UK.

R CORE TEAM (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/.

RAFTERY, A. E., LALIC, N. and GERLAND, P. (2014). Joint probabilistic projection of female and male life expectancy. *Demogr. Res.* **30** 795–822.

RAFTERY, A. E., LI, N., ŠEVČÍKOVÁ, H., GERLAND, P. and HEILIG, G. K. (2012). Bayesian probabilistic population projections for all countries. *Proc. Natl. Acad. Sci. USA* **109** 13915–13921.

RAFTERY, A. E., CHUNN, J. L., GERLAND, P. and ŠEVČÍKOVÁ, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography* **50** 777–801.

RENSHAW, A. E. and HABERMAN, S. (2003). Lee–Carter mortality forecasting with age-specific enhancement. *Insurance Math. Econom.* **33** 255–272. MR2039286

RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **53** 233–243. MR1094283

ŠEVČÍKOVÁ, H. and RAFTERY, A. (2015). bayesLife: Bayesian Projection of Life Expectancy. R package version 2.2-0. Available at http://CRAN.R-project.org/package=bayesLife.

ŠEVČÍKOVÁ, H., LI, N., KANTOROVÁ, V., GERLAND, P. and RAFTERY, A. E. (2015). Age-specific mortality and fertility rates for probabilistic population projections. Univ. Washington. Working paper. Available at http://arxiv.org/abs/1503.05215.

SHANG, H. L. (2016). Supplement to "Mortality and life expectancy forecasting for a group of populations in developed countries: a multilevel functional data method." DOI:10.1214/16-AOAS953SUPP.

SHANG, H. L., BOOTH, H. and HYNDMAN, R. J. (2011). Point and interval forecasts of mortality rates and life expectancy: A comparison of ten principal component methods. *Demogr. Res.* **25** 173–214.

SHANG, H. L. and HYNDMAN, R. J. (2016). Grouped functional time series forecasting: An application to age-specific mortality rates. Monash Univ. Working paper 04/16. Available at http://business.monash.edu/econometrics-and-business-statistics/research/publications/ebs/wp04-16.pdf.

TICKLE, L. and BOOTH, H. (2014). The longevity prospects of Australian seniors: An evaluation of forecast method and outcome. *Asia-Pacific Journal of Risk and Insurance* **8** 259–292.

WIŚNIOWSKI, A., SMITH, P. W. F., BIJAK, J., RAYMER, J. and FORSTER, J. (2015). Bayesian population forecasting: Extending the Lee–Carter method. *Demography* **52** 1035–1059.

WOODS, C. and DUNSTAN, K. (2014). Forecasting mortality in New Zealand. Working paper 14-01, Statistics New Zealand. Available at http://www.stats.govt.nz/methods/research-papers/working-papers-original/forecasting-mortality-14-01.aspx.

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

ZHANG, J.-T. (2014). *Analysis of Variance for Functional Data. Monographs on Statistics and Applied Probability* **127**. CRC Press, Boca Raton, FL. MR3185072

ZIVOT, E. and WANG, J. (2003). *Modeling Financial Time Series with S-Plus®*. Springer, New York. MR2000944

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1673–1698
DOI: 10.1214/16-AOAS954

DATA MINING TO INVESTIGATE THE METEOROLOGICAL DRIVERS FOR EXTREME GROUND LEVEL OZONE EVENTS[1]

BY BROOK T. RUSSELL*, DANIEL S. COOLEY†,2, WILLIAM C. PORTER‡, BRIAN J. REICH§ AND COLETTE L. HEALD‡

Clemson University, Colorado State University†, Massachusetts Institute of Technology‡ and North Carolina State University§*

This project aims to explore which combinations of meteorological conditions are associated with extreme ground level ozone conditions. Our approach focuses only on the tail by optimizing the tail dependence between the ozone response and functions of meteorological covariates. Since there is a long list of possible meteorological covariates, the space of possible models cannot be explored completely. Consequently, we perform data mining within the model selection context, employing an automated model search procedure. Our study is unique among extremes applications, as optimizing tail dependence has not previously been attempted, and it presents new challenges, such as requiring a smooth threshold. We present a simulation study which shows that the method can detect complicated conditions leading to extreme responses and resists overfitting. We apply the method to ozone data for Atlanta and Charlotte and find similar meteorological drivers for these two Southeastern US cities. We identify several covariates which help to differentiate the meteorological conditions which lead to extreme ozone levels from those which lead to merely high levels.

REFERENCES

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

BÉLISLE, C. J. P. (1992). Convergence theorems for a class of simulated annealing algorithms on $\mathbf{R}^d$. *J. Appl. Probab.* **29** 885–895. MR1188544

BELL, M. L., MCDERMOTT, A., ZEGER, S. L., SAMET, J. M. and DOMINICI, F. (2004). Ozone and short-term mortality in 95 US urban communities, 1987–2000. *JAMA J. Am. Med. Assoc.* **292** 2372–2378.

CHAUDHURI, S. and SOLAR-LEZAMA, A. (2011). Smoothing a program soundly and robustly. In *Computer Aided Verification*. *Lecture Notes in Computer Science* **6806** 277–292. Springer, Heidelberg. MR2870757

COLES, S. G., HEFFERNAN, J. and TAWN, J. (1999). Dependence measures for extreme value analysis. *Extremes* **2** 339–365.

COMPUTATIONAL AND INFORMATION SYSTEMS LABORATORY (2012). Yellowstone: IBM iDataPlex System (University Community Computing). Boulder, CO: National Center for Atmospheric Research. http://n2t.net/ark:/85065/d7wd3xhc.

Key words and phrases. Tail dependence, multivariate regular variation, constrained optimization, cross-validation, smooth threshold.

COOLEY, D., NAVEAU, P. and PONCET, P. (2006). *Dependence in Probability and Statistics*. *Lecture Notes in Statistics* **187** (P. BERTAIL, P. DOUKHAN and P. SOULIER, eds.). Springer, New York. MR2269087

DAVIS, R. A. and MIKOSCH, T. (2009). The extremogram: A correlogram for extreme events. *Bernoulli* **15** 977–1009. MR2597580

EASTOE, E. F. (2009). A hierarchical model for non-stationary multivariate extremes: A case study of surface-level ozone and NO_x data in the UK. *Environmetrics* **20** 428–444. MR2834808

EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. CRC press, Boca Raton.

EPA (2006). Air Quality Criteria for Ozone and Related Photochemical Oxidants (Final). U.S. Environmental Protection Agency Washington, DC, EPA/600/R-05/004aF-cF.

JACOB, D. J. and WINNER, D. A. (2009). Effect of climate change on air quality. *Atmos. Environ.* **43** 51–63.

KIRKPATRICK, S., GELATT, C. D. JR. and VECCHI, M. P. (1983). Optimization by simulated annealing. *Science* **220** 671–680. MR0702485

LARSSON, M. and RESNICK, S. I. (2012). Extremal dependence measure and extremogram: The regularly varying case. *Extremes* **15** 231–256. MR2915582

LEDFORD, A. W. and TAWN, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika* **83** 169–187. MR1399163

MARAUN, D., OSBORN, T. J. and RUST, H. W. (2011). The influence of synoptic airflow on UK daily precipitation extremes. Part I: Observed spatio-temporal relationships. *Clim. Dyn.* **36** 261–275.

MESINGER, F., DIMEGO, G., KALNAY, E., MITCHELL, K., SHAFRAN, P. C., EBISUZAKI, W., JOVIC, D., WOOLLEN, J., ROGERS, E., BERBERY, E. H. et al. (2006). North American Regional Reanalysis. *Bull. Am. Meteorol. Soc.* **87** 343–360.

MULLEN, K., ARDIA, D., GIL, D., WINDOVER, D. and CLINE, J. (2011). DEoptim: An R package for global optimization by differential evolution. *J. Stat. Softw.* **40** 1–26.

R DEVELOPMENT CORE TEAM (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

REICH, B., COOLEY, D., FOLEY, K., NAPELENOK, S. and SHABY, B. (2013). Extreme value analysis for evaluating ozone control strategies. *Ann. Appl. Stat.* **7** 739–762. MR3112916

RESNICK, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes*. *Applied Probability. a Series of the Applied Probability Trust* **4**. Springer, New York. MR0900810

RESNICK, S. (2004). The extremal dependence measure and asymptotic independence. *Stoch. Models* **20** 205–227. MR2048253

RESNICK, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, New York. MR2271424

RUSSELL, B. T., COOLEY, D. S., PORTER, W. C., REICH, B. J. and HEALD, C. L. (2016). Supplement to "Data mining to investigate the meteorological drivers for extreme ground level ozone events." DOI:10.1214/16-AOAS954SUPP.

SILLMANN, J., CROCI-MASPOLI, M., KALLACHE, M. and KATZ, R. W. (2011). Extreme cold winter temperatures in Europe under the influence of North Atlantic atmospheric blocking. *J. Climate* **24** 5899–5913.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VARADHAN, R. (2011). alabama: Constrained nonlinear optimization. R package version 2011.9-1.

WILSON, A., RAPPOLD, A. G., NEAS, L. M. and REICH, B. J. (2014). Modeling the effect of temperature on ozone-related mortality. *Ann. Appl. Stat.* **8** 1728–1749. MR3271351

YANG XIANG, GUBIAN, S., SUOMELA, B. and HOENG, J. (2013). Generalized simulated annealing for global optimization: The GenSA package. *The R Journal* **5** 13–29.

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1699–1724
DOI: 10.1214/16-AOAS956

MULTIPLE TESTING UNDER DEPENDENCE VIA GRAPHICAL MODELS

BY JIE LIU[1], CHUNMING ZHANG[2] AND DAVID PAGE[1]

University of Wisconsin-Madison

Large-scale multiple testing tasks often exhibit dependence. Leveraging the dependence between individual tests is still one challenging and important problem in statistics. With recent advances in graphical models, it is feasible to use them to capture the dependence among multiple hypotheses. We propose a multiple testing procedure which is based on a Markov-random-field-coupled mixture model. The underlying true states of hypotheses are represented by a latent binary Markov random field, and the observed test statistics appear as the coupled mixture variables. The model can be learned by a novel EM algorithm. The next step is to infer the posterior probability that each hypothesis is null (termed *local index of significance*), and the false discovery rate can be controlled accordingly. We also provide a semiparametric variation of the graphical model which is useful in the situation where f_1 (the density function of the test statistic under the alternative hypothesis) is heterogeneous among multiple hypotheses. This semiparametric approach exactly generalizes the local FDR procedure [*J. Amer. Statist. Assoc.* **96** (2001) 1151–1160] and connects with the BH procedure [*J. Roy. Statist. Soc. Ser. B* **57** (1995) 289–300]. Simulations show that the numerical performance of multiple testing can be improved substantially by using our procedure. We apply the procedure to a real-world genome-wide association study on breast cancer, and we identify several SNPs with strong association evidence.

REFERENCES

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

BENJAMINI, Y. and HELLER, R. (2007). False discovery rates for spatial signals. *J. Amer. Statist. Assoc.* **102** 1272–1281. MR2412549

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BENJAMINI, Y. and HOCHBERG, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* **25** 60–83.

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

BESAG, J. (1975). Statistical analysis of non-lattice data. *J. R. Stat. Soc., Ser. D Stat.* **24** 179–195.

BLANCHARD, G. and ROQUAIN, É. (2009). Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.* **10** 2837–2871. MR2579914

Key words and phrases. Multiple testing under dependence, graphical models, Markov random field, local index of significance, genome-wide association study.

CELEUX, G., FORBES, F. and PEYRARD, N. (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recogn.* **36** 131–144.

CROUSE, M. S., NOWAK, R. D. and BARANIUK, R. G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Process.* **46** 886–902. MR1665651

EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302

EFRON, B. (2010). *Large-Scale Inference. Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758

EFRON, B. and TIBSHIRANI, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **23** 70–86.

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571

EPANECHNIKOV, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory Probab. Appl.* **14** 153–158.

FAN, J., HAN, X. and GU, W. (2012). Control of the false discovery rate under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1045.

FARCOMENI, A. (2007). Some results on the control of the false discovery rate under dependence. *Scand. J. Stat.* **34** 275–297. MR2346640

FINNER, H. and ROTERS, M. (2002). Multiple hypotheses testing and expected number of type I errors. *Ann. Statist.* **30** 220–238. MR1892662

FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

GENOVESE, C. R., ROEDER, K. and WASSERMAN, L. (2006). False discovery control with p-value weighting. *Biometrika* **93** 509–524. MR2261439

GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 499–517. MR1924303

GENOVESE, C. and WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32** 1035–1061. MR2065197

GUEDJ, M., ROBIN, S., CELISSE, A. and NUEL, G. (2009). Kerfdr: A semi-parametric kernel-based approach to local false discovery rate estimation. *BMC Bioinformatics* **10** 84.

HINTON, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14** 1771–1800.

HUNTER, D. J., KRAFT, P., JACOBS, K. B., COX, D. G., YEAGER, M., HANKINSON, S. E., WACHOLDER, S., WANG, Z., WELCH, R., HUTCHINSON, A., WANG, J., YU, K., CHATTERJEE, N., ORR, N., WILLETT, W. C., COLDITZ, G. A., ZIEGLER, R. G., BERG, C. D., BUYS, S. S., MCCARTY, C. A., FEIGELSON, H. S., CALLE, E. E., THUN, M. J., HAYES, R. B., TUCKER, M., GERHARD, D. S., FRAUMENI, J. F., HOOVER, R. N., THOMAS, G. and CHANOCK, S. J. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39** 870–874.

INTERNATIONAL HAPMAP CONSORTIUM (2003). The international HapMap project. *Nature* **426** 789–796.

JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.

KIM, D. and ZHANG, C. (2014). Adaptive linear step-up multiple testing procedure with the bias-reduced estimator. *Statist. Probab. Lett.* **87** 31–39. MR3168932

KSCHISCHANG, F. R., FREY, B. J. and LOELIGER, H.-A. (2001). Factor graphs and the sum-product algorithm. *IEEE Trans. Inform. Theory* **47** 498–519. MR1820474

LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Roy. Statist. Soc. Ser. B* **50** 157–224. MR0964177

LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105** 18718–18723.

LIANG, K. and NETTLETON, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 163–182. MR2885844

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A. et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.

MCCARTY, C. A., WILKE, R. A., GIAMPIETRO, P. F., WESBROOK, S. D. and CALDWELL, M. D. (2005). Marshfield clinic personalized medicine research project (PMRP): Design, methods and recruitment for a large population-based biobank. *Personalized Medicine* **2** 49–79.

MCCARTY, C. A., CHISHOLM, R. L., CHUTE, C. G., KULLO, I. J., JARVIK, G. P., LARSON, E. B., LI, R., MASYS, D. R., RITCHIE, M. D., RODEN, D. M. et al. (2011). The eMERGE network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Medical Genomics* **4** 13.

MURPHY, K. P., WEISS, Y. and JORDAN, M. I. (1999). Loopy belief propagation for approximate inference: An empirical study. In *UAI* 467–475.

NGUYEN, V. H. and MATIAS, C. (2014). Nonparametric estimation of the density of the alternative hypothesis in a multiple testing setup. Application to local false discovery rate estimation. *ESAIM Probab. Stat.* **18** 584–612. MR3334005

OWEN, A. B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 411–426. MR2155346

ROBBINS, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1950 131–148. Univ. California Press, Berkeley. MR0044803

ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. i* 157–163. Univ. California Press, Berkeley. MR0084919

ROBIN, S., BAR-HEN, A., DAUDIN, J.-J. and PIERRE, L. (2007). A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Comput. Statist. Data Anal.* **51** 5483–5493. MR2407654

ROMANO, J. P., SHAIKH, A. M. and WOLF, M. (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST* **17** 417–442. MR2470085

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.* **27** 832–837. MR0079873

SARKAR, S. K. (2006). False discovery and false nondiscovery rates in single-step multiple testing procedures. *Ann. Statist.* **34** 394–415. MR2275247

SATROM, P., BIESINGER, J., LI, S. M., SMITH, D., THOMAS, L. F., MAJZOUB, K., RIVAS, G. E., ALLUIN, J., ROSSI, J. J., KRONTIRIS, T. G., WEITZEL, J., DALY, M. B., BENSON, A. B., KIRKWOOD, J. M., ODWYER, P. J., SUTPHEN, R., STEWART, J. A., JOHNSON, D. and LARSON, G. P. (2009). A risk variant in an miR-125b binding site in BMPR1B is associated with breast cancer pathogenesis. *Cancer Res.* **69** 7459–7465.

SCHRAUDOLPH, N. N. (2010). Polynomial-time exact inference in NP-hard binary MRFs via reweighted perfect matching. In *AISTATS*.

SCHRAUDOLPH, N. N. and KAMENETSKY, D. (2009). Efficient exact inference in planar Ising models. In *NIPS*.

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. MR1924302

STOREY, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035. MR2036398

STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 187–205. MR2035766

SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657

SUN, W. and CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 393–424. MR2649603

SVENSON, U., NORDFJÄLL, K., STEGMAYR, B., MANJER, J., NILSSON, P., TAVELIN, B., HENRIKSSON, R., LENNER, P. and ROOS, G. (2008). Breast cancer survival is associated with telomere length in peripheral blood cells. *Cancer Res.* **68** 3618–3623.

TIELEMAN, T. (2008). Training restricted Boltzmann machines using approximations to the likelihood gradient. In *ICML* 1064–1071.

WAINWRIGHT, M. J., JAAKKOLA, T. S. and WILLSKY, A. S. (2003a). Tree-reweighted belief propagation algorithms and approximate ML estimation via pseudo-moment matching. In *AISTATS*.

WAINWRIGHT, M. J., JAAKKOLA, T. S. and WILLSKY, A. S. (2003b). Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Trans. Inform. Theory* **49** 1120–1146. MR1984817

WEI, Z., SUN, W., WANG, K. and HAKONARSON, H. (2009). Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics* **25** 2802–2808.

WEISS, Y. (2000). Correctness of local probability propagation in graphical models with loops. *Neural Comput.* **12** 1–41.

WELLING, M. and SUTTON, C. (2005). Learning in Markov random fields with contrastive free energies. In *AISTATS*.

WU, W. B. (2008). On false discovery control under dependence. *Ann. Statist.* **36** 364–380. MR2387975

XIAO, J., ZHU, W. and GUO, J. (2013). Large-scale multiple testing in genome-wide association studies via region-specific hidden Markov models. *BMC Bioinformatics* **14** 282.

YEDIDIA, J. S., FREEMAN, W. T. and WEISS, Y. (2000). Generalized belief propagation. In *NIPS* 689–695. MIT Press, Cambridge.

YEKUTIELI, D. and BENJAMINI, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Statist. Plann. Inference* **82** 171–196. MR1736442

ZHANG, Y., BRADY, M. and SMITH, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imag.* **20** 45–57.

ZHANG, C., FAN, J. and YU, T. (2011). Multiple testing via FDR_L for large-scale imaging data. *Ann. Statist.* **39** 613–642. MR2797858

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1725–1756
DOI: 10.1214/16-AOAS957

SPATIALLY INHOMOGENEOUS BACKGROUND RATE ESTIMATORS AND UNCERTAINTY QUANTIFICATION FOR NONPARAMETRIC HAWKES POINT PROCESS MODELS OF EARTHQUAKE OCCURRENCES

BY ERIC WARREN FOX, FREDERIC PAIK SCHOENBERG AND JOSHUA SETH GORDON

University of California, Los Angeles

Space–time Hawkes point process models for the conditional rate of earthquake occurrences traditionally make many parametric assumptions about the form of the triggering function for the rate of aftershocks following an earthquake. As an alternative, Marsan and Lengliné [*Science* **319** (2008) 1076–1079] developed a completely nonparametric method that provides an estimate of a homogeneous background rate for mainshocks, and a histogram estimate of the triggering function. At each step of the procedure the model estimates rely on computing the probability each earthquake is a mainshock or aftershock of a previous event. The focus of this paper is the improvement and assessment of Marsan and Lengliné's method in the following ways: (a) the proposal of novel ways to incorporate a spatially inhomogeneous background rate; (b) adding error bars to the histogram estimates which quantify the sampling variability in the estimation of the underlying seismic process. A simulation study is designed to evaluate and validate the ability of our methods to recover the triggering function and spatially varying background rate. An application to earthquake data from the Tohoku District in Japan is discussed at the end, and the results are compared to a well-established parametric model of seismicity for this region.

REFERENCES

ADELFIO, G. and CHIODI, M. (2013). Mixed estimation technique in semi-parametric space–time point processes for earthquake description. In 28*th International Workshop on Statistical Modeling* **1** 65–70. Palermo, Italy.

ADELFIO, G. and CHIODI, M. (2015). Alternated estimation in semi-parametric space–time branching-type point processes with application to seismic catalogs. *Stochastic Environmental Research and Risk Assessment* **29** 443–450.

BRILLINGER, D. R. (1998). Some wavelet analyses of point process data. In *The Thirty-First Asilomar Conference on Signals, Systems and Computers* **2** 1087–1091. IEEE Computer Society, Los Alamitos, CA.

DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes. Vol. I: Elementary Theory and Methods*, 2nd ed. Springer, New York. MR1950431

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application. Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge Univ. Press, Cambridge. MR1478673

Key words and phrases. Point processes, nonparametric estimation, Hawkes process, MISD, ETAS model, earthquake forecasting.

FOX, E. W., SCHOENBERG, F. P. and GORDON, J. S. (2016). Supplement to "Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences." DOI:10.1214/16-AOAS957SUPP.

GUTENBERG, B. and RICHTER, C. F. (1944). Frequency of earthquakes in California. *Bull. Seismol. Soc. Amer.* **34** 185–188.

HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. MR0278410

LENGLINÉ, O., ENESCU, B., PENG, Z. and SHIOMI, K. (2012). Decay and expansion of the early aftershock activity following the 2011, Mw9.0 Tohoku earthquake. *Geophysical Research Letters* **39** L18309.

LEWIS, P. A. W. and SHEDLER, G. S. (1979). Simulation of nonhomogeneous Poisson processes by thinning. *Naval Res. Logist. Quart.* **26** 403–413. MR0546120

MARSAN, D. and LENGLINÉ, O. (2008). Extending earthquakes' reach through cascading. *Science* **319** 1076–1079.

MARSAN, D. and LENGLINÉ, O. (2010). A new estimation of the decay of aftershock density with distance to the mainshock. *Journal of Geophysical Research*: *Solid Earth* **115** B09302.

MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. MR2816705

MUSMECI, F. and VERE-JONES, D. (1992). A space–time clustering model for historical earthquakes. *Ann. Inst. Statist. Math.* **44** 1–11.

NICHOLS, K. and SCHOENBERG, F. P. (2014). Assessing the dependency between the magnitudes of earthquakes and the magnitudes of their aftershocks. *Environmetrics* **25** 143–151. MR3200305

OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.

OGATA, Y. (1998). Space–time point-process models for earthquake occurrences. *Ann. Inst. Statist. Math.* **50** 379–402.

OGATA, Y. and KATSURA, K. (1988). Likelihood analysis of spatial inhomogeneity for marked point patterns. *Ann. Inst. Statist. Math.* **40** 29–39. MR0946013

SCHOENBERG, F. P. (2013). Facilitated estimation of ETAS. *Bull. Seismol. Soc. Amer.* **103** 601–605.

UTSU, T., OGATA, Y. and MATSU'URA, R. S. (1995). The centenary of the Omori formula for a decay law of aftershock activity. *Journal of Physics of the Earth* **43** 1–33.

VEEN, A. and SCHOENBERG, F. P. (2008). Estimation of space–time branching process models in seismology using an EM-type algorithm. *J. Amer. Statist. Assoc.* **103** 614–624. MR2523998

ZHUANG, J., OGATA, Y. and VERE-JONES, D. (2002). Stochastic declustering of space–time earthquake occurrences. *J. Amer. Statist. Assoc.* **97** 369–380. MR1941459

ZHUANG, J., OGATA, Y. and VERE-JONES, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research*: *Solid Earth* **109** B05301.

The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1757–1775
DOI: 10.1214/16-AOAS958

CORRECTION OF BIFURCATED RIVER FLOW MEASUREMENTS FROM HISTORICAL DATA: PAVING THE WAY FOR THE TEESTA WATER SHARING TREATY

BY KAUSHIK JANA*, DEBASIS SENGUPTA* AND KALYAN RUDRA†

*Indian Statistical Institute** *and West Bengal Pollution Control Board*†

In this paper, we consider an estimation problem arising in the measurement of bifurcated flow of the Teesta, a trans-boundary river flowing through India and Bangladesh. The location of measurement is an Indian Barrage, where a part of the flow is diverted from the main stream to a canal. The flows through the two channels are regulated by different control structures and are measured indirectly from the height of the water level and the dimensions of the control structures. The computational formula for the measurement involves a hydrological constant used as a multiplier. Empirical findings indicate that incorrect multipliers are currently used in the computational formula for the two channels. For implementing any water sharing treaty between the two countries, the measurements need to be brought to a common scale. For this purpose, we present a model with carefully considered assumptions to estimate the correction factor. The model permits diagnostic tests for validation of the assumptions. We provide a nonparametric and consistent estimator of the desired factor.

Analysis of historical flow data shows that a main stream flow measured as 100 cumec would be measured as 76 cumec if it is diverted through the canal. Adjustment of emerging measurements through this finding would help the governments of India and Bangladesh to effectively implement and monitor any water sharing agreement.

REFERENCES

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

CHEONG, W. F., PRAHL, S. A. and WELCH, A. J. (1990). A review of the optical properties of biological tissues. *IEEE J. Quantum Electron.* **26** 2166–2185.

COMON, P. (1994). Independent component analysis, a new concept? *Signal Process.* **36** 287–314.

DUNCLE, R. V. and BEVANS, J. T. (1956). An approximate analysis of the solar reflectance and tranmittance of a snow cover. *J. Meteor.* **13** 212–216.

GREENWELL, B. M. (2014). *Topics in Statistical Calibration*. Ph.D. thesis, Air Force Institute of Technology. MR3218115

HÁJEK, J., ŠIDÁK, Z. and SEN, P. K. (1999). *Theory of Rank Tests*, 2nd ed. *Probability and Mathematical Statistics*. Academic Press, San Diego, CA. MR1680991

HOLLANDER, M. and WOLFE, D. A. (1999). *Nonparametric Statistical Methods*, 2nd ed. Wiley, New York. MR1666064

Key words and phrases. Bifurcation, dependence measure, hydrological constant, independence, multiplicative distortion, trans-boundary river.

INDIA TODAY (2015). Mamata Banerjee raises Teesta issue with Sheikh Hasina, assures a breakthrough. *India Today*, Feb 25.

JAIN, S. C. (2001). *Open-Channel Flow*. Wiley, New York.

JANA, K., SENGUPTA, D. and RUDRA, K. (2016). Supplement to "Correction of bifurcated river flow measurements from historical data: Paving the way for the Teesta water sharing treaty." DOI:10.1214/16-AOAS958SUPP.

JHA, R. K. (2015). India-Bangladesh politics over Teesta river water sharing. *South Asia Monitor*, Jan 27.

KENDALL, M. G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–93.

LEVY, B. C. (2008). *Principles of Signal Detection and Parameter Estimation*, 1st ed. Springer, New York.

MENON, M. S. (2015). Time to look at Teesta. *The Indian Express*, Aug 13.

MILLER, R. G. JR. (1981). *Simultaneous Statistical Inference*, 2nd ed. Springer, New York. MR0612319

MOUDGIL, M. (2015). South Asian water wars: An improbability. *World Policy Insti.*, Sep 14. Available at http://www.worldpolicy.org/blog/2015/09/14/south-asian-water-wars-improbability.

OSBORNE, C. (1991). Statistical calibration: A review. *Int. Stat. Rev.* **59** 309–336.

PATTERSON, M. S., CHANCE, B. and WILSON, B. C. (1989). Time resolved reflectance and transmittance for the non-invasive measurement of tissue optical properties. *Appl. Opt.* **28** 2331–2336.

RUDRA, K. (2012). *Atlas of Changing River Courses in West Bengal*. Sea Explorers' Institute, Kolkata.

SENGUPTA, D. and JAMMALAMADAKA, S. R. (2003). *Linear Models*: *An Integrated Approach*. *Series on Multivariate Analysis* **6**. World Scientific, River Edge, NJ. MR1993512

SOLOMON, S. (2010). *Water*: *The Epic Struggle for Wealth*, *Power and Civilization*. Harper Collins, New York.

SPEARMAN, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* **15** 72–101.

SUBRAMANYA, K. (2013). *Engineering Hydrology*, 4th ed. Tata McGraw Hill Education, New Delhi.

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665

TASSAN, S. and FERRARI, G. M. (2002). A sensitivity analysis of the transmittance–reflectance method for measuring light absorption by aquatic particles. *J. Plankton Res.* **24** 757–774.

VIDAL, J. (2010). How water raises the political temperature between countries. *The Guardian*, June 25.

WILSON, B. C. and PATTERSON, M. S. (2008). The physics, biophysics and technology of photodynamic therapy. *Phys. Med. Biol.* **53** 61–106.

YU, X., HU, X. and XU, J. (2014). *Blind Source Separation*: *Theory and Applications*, 1st ed. Wiley, New York.